A FRAMEWORK FOR EXPLOITING MODULATION SPECTRAL FEATURES IN MUSIC DATA MINING AND OTHER APPLICATIONS

A Dissertation
Presented to
The Academic Faculty

By

Nashlie H. Sephus

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
August 2014

A FRAMEWORK FOR EXPLOITING MODULATION SPECTRAL FEATURES IN MUSIC DATA MINING AND OTHER APPLICATIONS

Approved by:

Dr. Aaron D. Lanterman, Advisor
Associate Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. William D. Hunt
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. David V. Anderson, Co-Advisor
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Ayanna M. Howard
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Mark A. Clements
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Alexander G. Lerch
Assistant Professor, School of Music
Georgia Institute of Technology

Date Approved: May 6, 2014

To my Lord and Savior, Jesus Christ, for bringing me thus far, and my family and friends whom have been so encouraging and supportive of me over the years. "I can do all things through Christ which strengtheneth me."-Philippians 4:13

ACKNOWLEDGMENTS

I thank my Lord and Savior, Jesus Christ, and all who have helped me through this process. My advisors, Dr. Aaron Lanterman and Dr. David Anderson, have pushed me to go the extra mile and to achieve new heights in my research. I thank them and my committee members, Dr. Mark Clements, Dr. William Hunt, Dr. Ayanna Howard, and Dr. Alexander Lerch, for accepting the challenge as we joined forces in this research. I also acknowledge the assistance and counsel from Dr. Chin Lee, Dr. Les Atlas (University of Washington), Dr. Pascal Clark (University of Washington), Dr. Shigeki Sagayama (University of Tokyo–along with PhD student Hideyuki Tachibana), and Dr. Dan P.W. Ellis (Columbia University).

My mentors, especially Dr. Donna Reese (Mississippi State University), Dr. Otis Smart (Emory University), and Mr. Michael Parran, both within and outside of Georgia Tech have motivated me to press on and have guided me in several ways. I thank Dr. Gary May for his efforts to inspire undergraduates in engineering to do research, in particular for founding the Summer Undergraduate Program in Engineering Research at Berkeley (SUPERB). My family, especially my mother (Ms. Tonie M. Sephus) and grandmother (Ms. Betty J. McPherson); friends; teachers and professors of the past and present; church family; and fellow colleagues have provided me with many resources, expertise, and helpful advice. I acknowledge the various people and departments I have worked with at Georgia Tech, such as the School of Electrical and Computer Engineering staff, Jill Auerbach and Julie Ridings from the Opportunities Research Scholars (ORS) Program, the Office of Undergraduate Research, the Office of the Vice President for Institute of Diversity (VPID), the Black Graduate Student Association (BGSA), the Center for the Enhancement of Teaching and Learning (CETL), the Center for Education Integrating Science, Mathematics and Computing (CEISMC), the Sam Nunn Security Fellowship/McArthur Foundation, the Virtuous Organization of Women (VOW) Bible Study, and the College of Engineering. They have

helped to mold me and enrich my Georgia Tech experience. Lastly, I am truly grateful for the Georgia Tech Center for Women, Science, and Technology (WST) and the co-directors Dr. Carol Colatrella, Dr. Mary Frank Fox, and Dr. Mary Lynn Realff who have funded the majority of my PhD research and who have helped to make my dream a reality.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The term “modulation” brings diverse concepts to mind, such as the transmission of sound over radio waves via amplitude modulation (AM) or frequency modulation (FM), musicians modulating “keys” within a musical piece, or an audio engineer using special effects to synthesize a sound. The concept of “modulation frequency” may be most commonly known in AM and FM radio, where music and talk shows are broadcast “on air” via a transmitted carrier frequency that is either modulated in frequency or amplitude by the message, or modulating, signal. The radio receiver is tuned to the carrier frequency to receive the broadcast signal by demodulating it into its modulator and carrier parts. The frequency of the modulator is usually much smaller than the carrier frequency, and therefore, a slowly varying envelope, or modulator, is formed in the time domain for AM and in the frequency domain for FM. When a signal is decomposed into frequency bands, demodulated into modulator and carrier pairs, and portrayed in a carrier frequency-versus-modulator frequency domain, significant information may be automatically observed about the signal. We refer to this domain as the *modulation spectral domain*.

The *modulation spectrum* is referred to as a windowed Fourier transform across time that produces an acoustic frequency versus modulation frequency representation of a signal. Previously, frameworks incorporating the discrete short-time modulation transform (DSTMT) and modulation spectrum have been designed mostly for filtering of speech signals. This modulation spectral domain is rarely, if ever, discussed in typical signal processing courses today, and we believe its current associated tools and applications are somewhat limited. We seek to revisit this domain to uncover more intuition, develop new concepts to extend its capabilities, and increase its applications, especially in the area of music data mining.

A recent interest has risen in using *modulation spectral features*, which are features in the modulation spectral domain, for music data mining. The field of music data mining,

also known as music information retrieval (MIR), has been rapidly developing over the past decade or so [2]. One reason for this development is the aim to develop frameworks leveraging the particular characteristics of music signals instead of simply copying methods previously applied to its speech-centered predecessors, such as speech recognition, speech synthesis, and speaker identification. This research seeks to broaden the perspective and use of an existing modulation filterbank framework by exploiting modulation features well suited for music signals.

The objective of this thesis is to develop a framework for extracting modulation spectral features from music and other signals. The purpose of extracting features from these signals is to perform data mining tasks, such as unsupervised source identification, unsupervised source separation, and audio synthesis. More specifically, this research emphasizes the following: the usefulness of the DSTMT and the modulation spectrum for music data mining tasks; a new approach to unsupervised source identification using modulation spectral features; a new approach to unsupervised source separation; a newly introduced analysis of FM features in an "AM-dominated" modulation spectra; and other applications. The objective of the unsupervised identification method is to automatically identify distinct sources of varying modulation content, a process that is currently manual and requires prior information about sources. The objective of the unsupervised source separation is to blindly separate sources in periodic segments of signals with varying modulation content, or temporal patterns. When combined, our unsupervised source identification and source separation make up a modulation spectral feature framework that may be capable of other applications as well, such as vibrato analysis and modulation analysis of EEG seizure signals.

CHAPTER 1

INTRODUCTION

The term "modulation" brings diverse concepts to mind, such as the transmission of sound over radio waves via amplitude modulation (AM) or frequency modulation (FM), musicians modulating "keys" within a musical piece, or an audio engineer using special effects to synthesize a sound. The concept of "modulation frequency" may be most commonly known in AM and FM radio, where music and talk shows are broadcast "on air" via a transmitted carrier frequency that is either modulated in frequency or amplitude by the message, or modulating, signal. The radio receiver is tuned to the carrier frequency to receive the broadcast signal by demodulating it into its modulator and carrier parts. The frequency of the modulator is usually much smaller than the carrier frequency, and therefore, a slowly varying envelope, or modulator, is formed in the time domain for AM and in the frequency domain for FM. When a signal is decomposed into frequency bands, demodulated into modulator and carrier pairs, and portrayed in a carrier frequency-versus-modulator frequency domain, significant information may be automatically observed about the signal. We refer to this domain as the *modulation spectral domain*.

The *modulation spectrum* is referred to as a windowed Fourier transform across time that produces an acoustic frequency versus modulation frequency representation of a signal. Previously, frameworks incorporating the discrete short-time modulation transform (DSTMT) and modulation spectrum have been designed mostly for filtering of speech signals. This modulation spectral domain is rarely, if ever, discussed in typical signal processing courses today, and we believe its current associated tools and applications are somewhat limited. We seek to revisit this domain to uncover more intuition, develop new concepts to extend its capabilities, and increase its applications, especially in the area of music data mining.

A recent interest has risen in using *modulation spectral features*, which are features in

the modulation spectral domain, for music data mining. The field of music data mining, also known as music information retrieval (MIR), has been rapidly developing over the past decade or so [2]. One reason for this development is the aim to develop frameworks leveraging the particular characteristics of music signals instead of simply copying methods previously applied to its speech-centered predecessors, such as speech recognition, speech synthesis, and speaker identification. This research seeks to broaden the perspective and use of an existing modulation filterbank framework by exploiting modulation features well suited for music signals.

The objective of this thesis is to develop a framework for extracting modulation spectral features from music and other signals. The purpose of extracting features from these signals is to perform data mining tasks, such as unsupervised source identification, unsupervised source separation, and audio synthesis. More specifically, this research emphasizes the following: the usefulness of the DSTMT and the modulation spectrum for music data mining tasks; a new approach to unsupervised source identification using modulation spectral features; a new approach to unsupervised source separation; a newly introduced analysis of FM features in an "AM-dominated" modulation spectra; and other applications. The objective of the unsupervised identification method is to automatically identify distinct sources of varying modulation content, a process that is currently manual and requires prior information about sources. The objective of the unsupervised source separation is to blindly separate sources in periodic segments of signals with varying modulation content, or temporal patterns. When combined, our unsupervised source identification and source separation make up a modulation spectral feature framework that may be capable of other applications as well, such as vibrato analysis and modulation analysis of EEG seizure signals.

This introduction is followed by five chapters. Chapter 2 is a background on modulation features, state-of-the-art methods for such features, and relevant issues in the scope of this research. Here, we also demonstrate an application to unsupervised source identification.

Chapter 3 discusses FM features in the "AM-dominated" modulation spectra with some applications in vibrato analysis. In Chapter 4, the methodology of the framework for unsupervised source identification and source separation is discussed along with explanation of parameter usage, experimental results, and listening test results. Chapter 5 presents an additional framework application involving modulation analysis of seizure signals. Lastly, Chapter 6 summarizes the thesis research and discusses possible future work.

CHAPTER 2

BACKGROUND ON MODULATION SPECTRAL FEATURES WITH APPLICATION TO UNSUPERVISED SOURCE IDENTIFICATION IN MUSIC

[1]Modulation frequency analysis has been studied in several research areas, such as communications, filtering, coding of digital signals, and representations of neurons used for biomedical research. This research lead to the development of the modulation spectrum (an acoustic frequency versus modulation frequency plot of a time window from a signal) and filterbank framework for signal processing in speaker identification, audio coding, and synthesis [1]. The term "modulation spectral features" has been used rather loosely, but we refer to them as temporal patterns (whether long-term or short-term) that may be revealed from the modulation spectrum. This chapter reviews previous uses of general modulation features and encourages new applications in music data mining by focusing on the modulation spectral domain. We also further motivate the use of such features by presenting our preliminary, unsupervised identification algorithm for sources of varying temporal patterns, or rhythmic structure. The title is inspired by and parallels a 1997 conference publication by Greenberg and Kingsbury [4] called "The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech." In addition to the change of emphasis from speech to music, our change of "The Modulation Spectrogram" to "Modulation Spectral Features" and "an Invariant Representation" to "Invariant Representations" reflects out broader scope. Greenberg and Kingsbury focused primarily on the noise suppression properties of their modulation spectrogram compared with the traditional frequency-versus-time spectrogram. Our exposition covers wider ground and emphasizes that modulation spectral features may be appropriate in different musical contexts.

[1]This chapter is modified from *Sephus, N. H., Lanterman, A. D., & Anderson, D. V. (2014). Modulation Spectral Features: In Pursuit of Invariant Representations of Music with Application to Unsupervised Source Identification. Journal of New Music Research, To appear in the upcoming Special Issue on Music Rhythm.* [3].

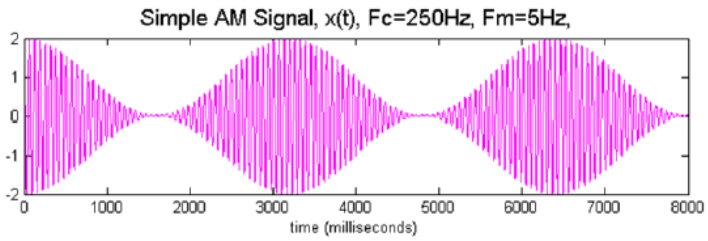


Figure 1: Example of a simple AM signal with a carrier frequency of 250 Hz (middle C) and a modulation frequency of 5 Hz.

2.1 Background

2.1.1 Motivation for Modulation Spectral Features

In music, determining differences in modulation between two different instruments may reveal a distinction between their timbre. The modulation frequency range for "tremolo" is defined as being 4-8 Hz and "roughness" is defined as being 8-10 Hz [5], while vibrato may be another range of frequencies. For complicated samples of speech and music with multiple instruments and multiple speakers, extracting carriers and modulators, or slow-varying envelope signals, associated with multiple signal components can become quite difficult. The modulation spectral domain may address such complicated signals.

2.1.2 A Modulation Filterbank Framework

We will begin by describing the basic concept of modulation frequency before discussing an existing modulation filterbank framework, which is the basis of our research. Consider a simple amplitude modulated signal $x(t)$ with the general form $x(t) = m(t)c(t)$, where $m(t)$ is a modulator, or slow-varying envelope, and $c(t)$ is a carrier signal with a much larger frequency for AM and $x(t) = c(t + Im(t))$ for FM, where I controls the amount of frequency variation. (FM is discussed in more detail in Chapter 3.) Consider a simple, sinusoidal AM signal defined as $x(t) = \{A_c + \cos(\omega_m t)\}\cos(\omega_c t)$, where A_c is the amplitude of the carrier signal, $\omega_m = 2\pi f_m$, f_m is the modulator frequency (Hz), $\omega_c = 2\pi f_c$, and f_c is the carrier frequency (Hz). An example is shown in Fig. 1.

The most well-known form of general decomposition involves filterbank analysis. In this case, the signal $x(t)$ is divided into k subbands $x_k(t) = x(t) * h_k(t)$, where the impulse

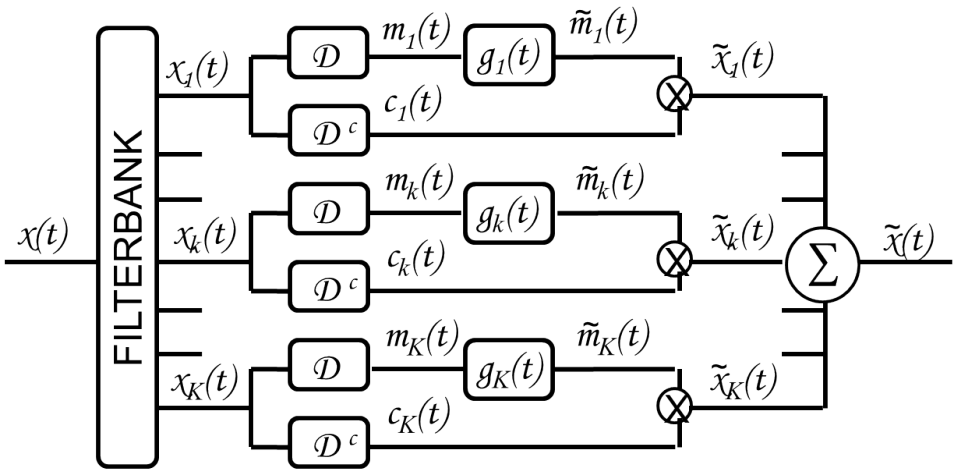


Figure 2: Diagram of an existing modulation filterbank system. [1]

responses of the bandpass filters are denoted as $h_k(t)$ and $*$ represents convolution. In an existing modulation filterbank framework (as shown in Fig. 2) [1], the resulting subband signals $x_k(t)$ are decomposed into a modulator $m_k(t)$ and a carrier $c_k(t)$ after undergoing modulator, or envelope, detection ($\mathfrak{D}$) and carrier estimation ($\mathfrak{D}^c$), as shown in Fig. 2 taken from [1]. Some type of modulation filter impulse response ($g_k(t)$) may be used to modify the modulator, such as modulation filtering, before multiplying with its associated carrier. The resulting components $\tilde{x}_k(t)$ are then summed to synthesize $\tilde{x}(t)$.

2.1.3 The Modulation Spectrum

For better visual comparison of signals with multiple modulations, the *modulation spectrogram* was introduced in 1997 [4]. The authors of [4] were partially influenced by RASTA [6], a modulation-inspired feature extraction method often used with speech and sometimes with other kinds of audio. Later, the *modulation spectrum*[2] was formalized into a standard theoretical definition [7]. The modulation spectrum is a visual representation of a signal's amplitude in terms of "modulation frequencies" (frequencies of the modulators), which are displayed on the x-axis, as it correlates to its "acoustic frequencies" (frequencies

[2]The "modulation spectrogram" developed in [4] differs from the "modulation spectrum" presented in [7] and used throughout this research, although both seek the same sort of two-dimensional acoustic-frequency-versus-modulation-frequency representation. We cite [4] to provide inspirational context.

of the carriers), which are displayed on the y-axis. Specifically, the modulation spectrum encodes the temporal variation of spectral energy in a signal by taking the modulation transform for a given length of the signal; either an entire signal or a shorter time window of the signal may be analyzed at once. This modulation transform, formally called the discrete short-time modulation transform (DSTMT), is defined as

$$X(l,i,k) = DSTMT\{x(n)\} \tag{1}$$

$$\triangleq DSTFT\{\mathfrak{D}_{\text{modulator}}\{DSTFT\{x(n)\}\}\} \tag{2}$$

$$= \sum_m \mathfrak{D}_{\text{modulator}}\{\sum_n x(n)w(m-n)e^{j2\pi nk/K}\}v(l-m)e^{j2\pi mi/I} \tag{3}$$

for $i = 0,\ldots,I-1$ and $k = 0,\ldots,K-1$, where $w(n)$ and $v(m)$ are analysis windows typically defined in the discrete short-time Fourier transforms (DSTFT) and $\mathfrak{D}_{\text{modulator}}$ is the method of detecting the modulators, or slow-varying envelopes, along carrier, or acoustic, frequency subbands (refer to the time-domain model in the left half of Fig. 2) [1]. The k variable indexes acoustic frequencies (typically displayed along the y-axis), the i variable indexes modulation frequencies (typically displayed along the x-axis), and the l variable indexes the time-window segment of the signal being analyzed (in the case that the entire signal is not analyzed at once). The Modulation Toolbox for MATLAB [8] includes functions for demodulating signals along subbands, plotting modulation spectra, filtering modulator signals, and other resourceful tools. Figure 3 shows what the simple AM equivalent signal's spectrogram and modulation spectrum would look like. This observation may be thought of in terms of short-time features for a single music note played by an instrument. A long-term example across a lengthy window for some repetitious, periodic segment of a song would be a source, such as a flute, that is played at middle C (about 262 Hz) every 0.2 secs (or a frequency period of 5 Hz).

2.1.4 Modulation Spectral Features

After computing the modulation spectrum of an analysis window from a signal, some features may be extracted to quantitatively summarize interesting aspects of the modulation

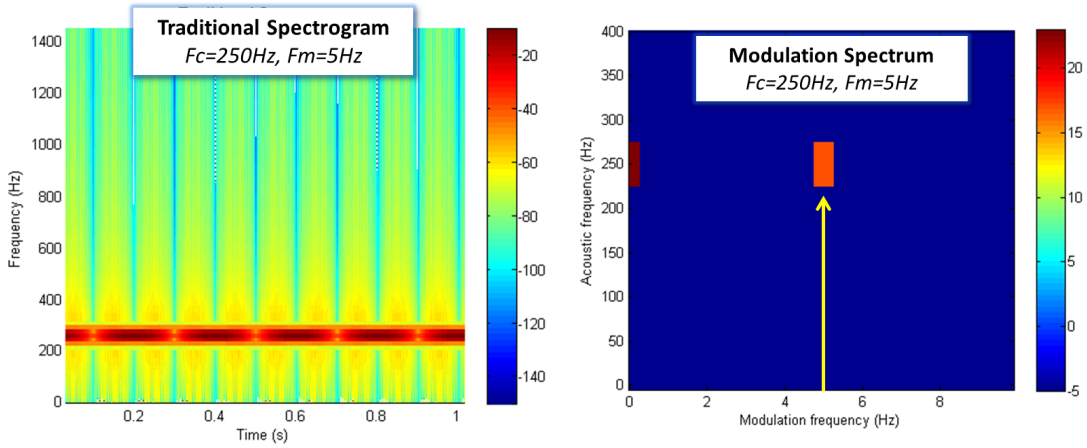


Figure 3: Example of a traditional spectrogram (left) and a modulation spectrum (right) containing a simple AM signal with a carrier frequency of 250 Hz (middle C) and a modulation frequency of 5Hz.

spectrum. In some cases, this is more efficient than comparing the entire modulation spectra directly and is a form of feature selection. For traditional spectral comparisons, features such as mel-frequency cepstral coefficients (MFCCs), spectral flux, spectral centroid, and spectral dissonance are commonly used [9]. Also, similar to the way the traditional frequency-versus-time spectrogram is treated as an image and visual features are extracted [10], [11], [12], the same can be done with the modulation spectrum. These modulation spectral features have been explored in data mining, initially for speech [13]. Recently, these explorations have been expanded to include music. The most notable features for modulation spectra would be the modulation spectral contrast (MSC), which was used for long-term signal analysis in a music genre classification study [14]. MSC is calculated as the difference between the modulation spectral peak (MSP) and the modulation spectral valley (MSV) of each logarithmically spaced modulation subband. Another feature called the cepstral modulation ratio has been used for hierarchical classification of content in audio signals [15] along with previous development of a feature set for speech derived from the modulation spectrum but on a mel-frequency scale [16]. For data mining in music, modulation spectral features have been combined with other features to increase robustness [17]. Considering the modulation feature frameworks for music, relatively few

exist for music signals (and are limited in a need for prior information about sources or limited in applicable types of signals). The majority of these frameworks have been focussed on the source separation and enhancement of speech signals.

2.1.5 Advantages of Modulation Features

2.1.5.1 Perceptual Aspects

One motivator for using modulation features for audio data mining is their relationship to the human auditory system [18], [19]. The human auditory system effectively divides sound into frequency bands, each of which is transduced into neural pulses by "hair cells" located in the inner ear. Although the complete processing chain is complicated, it is clear that the sound percept at each frequency is related to the short-term average firing rate of neurons. Thus, envelope extraction is performed on subbands of the audio. There is also evidence that further modulation analysis is performed in the auditory cortex [20]. Researchers found that modifying a signal's modulator directly affected the intelligibility of the signal, where as modifying the carrier did not have much of an effect on the signal's intelligibility [21]. Modulation features have been linked to several perceptual models for data mining associated with the auditory system, such as Perceptual Linear Predictive (PLP) analysis [22], multiple-band perceptual modulation analysis [23], temporal and modulation coherence in auditory scene analysis [24], timbre modeling [25], and signal texture modeling [26]. These and similar experiments illustrate that modulation features may be used for identifying content in acoustic signals in a way that mimics the human auditory system.

2.1.5.2 Longer-Term *Signal Representation*

Another useful property of modulation features is their ability to represent a signal's long-term temporal patterns, or possible models for rhythmic structure, when provided larger analysis windows. For instance, as shown in Section 2.1.3, one modulation spectrum computed for half a second of a signal may permit focus on the structure of a single note played by a single instrument, where as one modulation spectrum computed for several seconds

of a song may permit the analysis of a tempo or rhythm in a "bar" or "phrase" of music. MFCCs, which are one of the most common feature sets for classification in audio, are prominently used to compute features for short-time analysis windows of around 10 ms to 100 ms [27], [28], [29]. However, since MFCCs best extract information from small analysis windows and only within the lower, mel-frequency range [28], they may have limited applicability to a more "global" signal representation, which may be more desirable in some information retrieval tasks.

2.1.5.3 Noise Invariant Representation

In some classification tasks, modulation features have also outperformed MFCCs and other features in noisy environments [30], [31], [13]. Noise (as in random noise) that is evident in traditional frequency-versus-time spectrograms often tends to be less visible in modulation spectra [32] because noise tends to not possess a distinct modulation structure, or slow-varying envelope. In cases where unwanted noise has some modulation structure, such structure may differ from that of the sources of interest; hence, the noise will appear at a different location on the modulation spectrum making the features of the target sources easy to distinguish and identify.

2.1.5.4 Visible Characteristics in Modulation Spectra

In audio data mining, visual analysis has typically been done by comparing signals characterized by time-domain representations, traditional frequency spectra, and traditional frequency-versus-time spectrograms. In many cases, modulation spectra may more readily indicate interesting, distinguishing aspects. For example, simultaneous sounds with overlapping pitch in a given signal may be difficult to visibly distinguish in the traditional frequency-versus-time spectrogram, but this same signal may be non-overlapping in modulation content [7], [1]. If parameters are set appropriately, modulation spectra may be used to pick out a particular sound in a signal and modify that particular sound while leaving the rest of the signal mostly intact. Such modifications might include amplification, attenuation, or removal of a particular acoustic frequency's modulation content [33]. These

methods have been particularly useful in speech enhancement [34], [35], noise reduction, and source separation applications [36]. Overall, modulation spectra serve as a holistic form of modulation features for visibly comparing signals and sounds within signals.

2.1.6 Exploiting Modulation Features in Music Data Mining

Data mining for music signals can be challenging given the many possible variables and inconsistencies. For example, a musical piece may contain several parts, various styles, and many sources whether vocal, harmonic, or percussive. A continuing goal in music data mining research is the development of representations that are invariant to such complexities. We briefly note that general forms of modulation features have been used with data modeling and mining tools such as Gaussian Mixture Models (GMMs) [37], [38], [39], [40], Hidden Markov Models (HMMs) [41], and support vector machines (SVMs) [42], [43], [44], [45], [46], [47]. Decomposition and filtering tools such as Gabor filters [48], [49], [50], singular value decomposition (SVD) [51], [44], [45], and nonnegative matrix factorization (NMF) [52], [44] have also been explored with some forms of these features.

Recurring themes throughout the literature emphasize the main reason why modulation frequency features are ideal for music data mining [53]. Specifically, modulation content may be modeled independently of pitch [54]. Modulation features are useful for modeling either long-term or short-term time-varying information in music signals. Preprocessing stages, such as onset detection [55], may benefit from modulation features since quick, significant changes in slow-varying envelopes can indicate onsets. Other applications of modulation frequency features have included the representation of rhythmic features [56] and the extraction of periodic signals [57].

2.1.6.1 Genre Classification and Emotion Detection

Researchers in MIR have exploited the ability of *modulation spectral features* to represent long-term information. In 2007, the octave-based modulation spectral contrast (OMSC),

which represents such long-term behavior, was developed for genre classification [58]. One implementation of genre classification focused on using modulation spectral features and MFCCs to model the static and transitional portions of a song [59], [14]. Another feature set developed for genre classification uses a super-vector of features from modulation spectra, such as contrast, valley, energy, centroid, and flatness, along with others such as MFCCs and timbral features [17]. A framework was developed that consisted of a fusion of MFCCs and modulation spectra called highly-resolved cepstral modulation spectra [60]. Some success has also been obtained with classifiers based solely on modulation features (in contrast with the aforementioned hybrid techniques) that use sparse representations of detected modulation frequencies [61] and nonnegative principal component analysis (PCA) [44].

Emotion detection in music (also known as mood classification) is a popular and emerging area of study. The ability of modulation features to capture long-term characteristics of a signal makes them desirable for this application as well. One metric developed uses modulation feature extraction to detect the tempo of a song as a main indicator of emotion [62]. Other methods utilize the modulation spectrum directly for successful coding of emotion in music signals via OMSC-modified features called feature-based modulation flatness measures and feature-based modulation crest measures [63].

2.1.6.2 Timbre Modeling and Instrument Recognition

Timbre may be defined as the perceptual quality of sound beyond pitch and loudness [64], or timbre may be defined as a perception of sound which encompasses several distinguishable aspects, not necessarily excluding pitch and dynamics [65]. For example, two different instrument types when playing the same note, i.e., same pitch, likely will still sound different because of other characteristics that allude to its timbre. Modulation spectral features provide information about other characteristics of the notes played; they can be useful for modeling distinctions in timbre. As a commonality, timbre may be affected by characteristics such as duration, attack, harmonic structure, overall spectral shape, and decay [66] while often being described using words such as “brightness,” “fullness,” and

"activity" [67]. Since these characteristics and descriptions are most related to the temporal characteristics of a signal, modulation features may be well-equipped for these representations and, thus, have been used to encode timbre [68]. In polyphonic music, modulation features have been incorporated to model perceptual dimensions of timbre, which may be directly related to how humans classify genre, style, mood, and emotion in music [67].

Timbre modeling has lead to developments in automatic musical instrument recognition. AM and FM features have been defined for musical instrument recognition for isolated sounds [50]. In these experiments, when combining AM and FM features with MFCCs versus MFCCs alone, a 60 percent error rate reduction was achieved. Modulation analysis for timbre-modeling and classification methods overcomes commonly mistaken assumptions made in short-time analysis, such as statistical independence amongst analysis frames, which are less relevant when exploring longer-term features [69]. One real-time implementation of musical instrument recognition extracts features from the modulation spectrum and uses a pre-defined modulation structure and pitch analysis for each instrument [52].

In contrast, modulation features are less useful for modeling timbre when temporal patterns are not present, i.e. when little-to-no modulation content exists. For example, a trumpet played with vibrato will contain more modulation content than a flat trumpet note, and more modulation content leads to modulation features capable of modeling timbre more distinctively. Regardless, some modulation information when fused with other relevant features may be beneficial for automatically classifying musical instrument sounds [70].

2.1.6.3 Source Separation using Modulation Features

Source separation in the music domain consists of separating harmonic sounds from percussive sounds, vocals from musical instruments, particular instruments from other instruments, particular vocals from other vocals, melody from accompaniment, and music from noise. Some source separation methods applied to music have explored differences in traditional frequency-versus-time spectrograms. Other tasks using modulation features have

been employed to separate speech from music [71], speech from non-speech [42], [45], and to perform drum track extraction [72]. Also, recognition of sources using modulation features may lead to the development of semi-supervised source separation via modulation filtering, i.e. reducing or eliminating sources with a particular modulation frequency [36].

For separating widely contrasting timbres in monaural music signals, such as separating repeated flute and castanet notes, the modulation filtering method in [73] was successful in separation due to sources being significantly non-overlapping in pitch. However, prior information was needed to identify the sources, such as pitch information and knowing the number of sources present. Other methods combine various types of features, such as pitch and amplitude modulation, for resolving overlapping harmonics in source separation [74]. Another framework used spectral modulation for more informed drum source separation where the different drum components could be separated if their temporal patterns were unique [75]. For separation of vocals and music, a method was developed that decomposes audio signals across multiple acoustic bands (100 Hz to 1500 Hz being the most relevant) into AM and FM components using the nonlinear Teager-Kaiser energy operator [71]. In cases where the sounds were significantly overlapping in pitch and the modulation content was similar, the modulation features did not perform as well (as we would expect with any source separation algorithm). Separation can be difficult given that a useful modulator envelope may not always be recovered for synthesis, and the reconstructed signal may contain artifacts, which causes it to loose some intelligibility and quality.

2.2 Issues with Modulation Features for Data Mining

Some disadvantages of using modulation features (as they have been employed in existing frameworks) for music data mining tasks, as well as audio data mining in general, have been mentioned in the previous section. Some additional issues have been reported throughout the literature, which may have interfered with further progress on employing modulation features for music signals. In this section, we discuss some of these issues and possible

resolutions.

A major factor in developing modulation features involves the type of envelope detection employed [76]. When decomposing a signal with Hilbert demodulation using the modulation filterbank framework mentioned in Section 2.1.2, the assumption that AM/FM envelopes are real and nonnegative must be reasonable [36]. The real condition for natural signals is more likely; however, the nonnegative condition may not always be the case. For our experiments in forthcoming Section 2.3, we show that these assumptions are safe for music signals since they exhibit strong temporal patterns. Also, Hilbert demodulation is simpler in implementation, especially for complicated music signals that contain several modulator-carrier pairs. (More details about why we use Hilbert demodulation is discussed in Section 3.2.1).

On the other hand, in the case of non-stationary signals, such as speech or music that contains little-to-no temporal patterns, *coherent* demodulation methods [73] theoretically provide better estimates of the carrier signals, thus providing better estimates of the modulators. These methods use a technique that incorporates the instantaneous frequency along acoustic frequency subbands while bandlimiting the modulators. Regardless of which demodulation method is used, the act of modifying or synthesizing the modulator(s) after decomposing a signal and then attempting to reconstruct the signal may produce extraneous artifacts and reduce quality [77], [40], [78].

As mentioned in Section 2.1.2, decomposition of a signal is typically done via a filterbank prior to demodulation. Modulation features are highly sensitive to the choice of filterbank parameters and the length of the analysis window. Inappropriate parameter settings lead to issues with ensuring that the modulation spectrum is plotted at a proper resolution to display enough discriminating information in modulation frequency content. Since the parameters are dependent on the signal, the idea of forming a universal representation independent of the signal is discouraged. With shorter subband widths, modulation spectra can resolve nearby frequencies as distinct constant tones, while longer subband widths resolve

those same nearby frequencies as a single beating tone. Ideally, the widths of the acoustic frequency subbands must be wide enough to demodulate the full amplitude of an envelope signal, or modulator, from a particular sound source. For example, a harmonic instrument may have four harmonics that are mistakenly divided into either more than four or less than four frequency subbands; an ideal acoustic subband width for a filterbank would isolate each harmonic fully into its own separate subband.

In the modulation spectrum, modulation frequency resolution (along the x-axis) is also an important parameter. Sounds with strong temporal patterns that occur n times per second would need a resolution of at least $1/(2n)$ Hz per pixel. This resolution may be increased to cluster patterns into one pixel or decreased to show more detailed modulation content. For example, perhaps we are analyzing a region of time from a signal that contains both a trumpet note repeated once every second (1 Hz) and a trumpet note with the same pitch repeated once every 2 seconds (0.5 Hz). With a modulation frequency resolution of only 1 Hz per pixel along the x-axis, the modulation spectrum would show only one clustered component at 1 Hz, i.e. it would be unable to show the 0.5 Hz component due to the low resolution. In some cases, however, this low resolution may be favorable to isolate some targeted components that may occur on whole-integer modulation frequencies. In Section 2.3, we present preliminary results from our proposed unsupervised source identification algorithm that was motivated by our literature survey. In our preliminary experiments, we have identified the most useful parameters and resolution for modulation spectra to reveal useful feature information for music.

2.3 Unsupervised Source Identification with Modulation Spectral Features

Motivated by our literature review, we sought to exploit the benefits of modulation spectral features for music by developing a preliminary method for unsupervised source identification. The objective of the method is to automatically identify sources of various temporal

patterns, or distinct long-term modulation features. We further discuss the properties of such features and verify the method's capabilities to perform source identification without having prior knowledge about the number of sources present, the pitch of these sources, or their individual modulation characteristics. We conduct pilot-study experiments on music signals from commonly used music datasets (mentioned later), whether signals contain percussive sources, harmonic sources, or mixtures (with and without vocals). Also, we quantitatively evaluate the algorithm's performance by comparing similarity of the modulation features of the automatically identified sources with that of the original source.

We extend the concept of *longer-term* analysis windows so that temporal patterns are visible in the modulation spectrum, thus allowing us to automatically recognize the presence of sources in music signals. The proposed unsupervised source identification method is motivated by a similar method called harmonic-percussive sound separation (HPSS) [11]. Another similar technique finds rhythmic similarity of music by a method involving dynamic periodicity warping in [79]. Since modulation envelopes are consistent across any time window segments covering the same periodicity of a signal, modulation spectra are invariant to slight misalignments in time windows between signals. We note that this may also cause unwanted groupings of sources but establish threshold parameters of sensitivity to account for this preference.

2.3.1 Method

2.3.1.1 Properties of Modulation Spectral Features

The modulation spectrum more directly and visibly reveals information for our task than traditional frequency-vs-time spectrograms. Our algorithm involves finding where the dominant modulation frequencies are present horizontally along each row of the modulation spectrum and grouping harmonics that appear vertically along columns. We have defined properties, summarized below, for automatically identifying the presence of sources in music using these visual features from the modulation spectrum. Currently, we use the Hilbert demodulation method, which is well suited for our data mining tasks for music,

simpler to implement than coherent demodulation methods, faster in processing time [80], and relates closely to perceptual mapping. The properties of modulation spectral features, as demonstrated in Fig. 4, are as follows:

- Modulation spectral features of a source with strong temporal patterns will appear at multiples of its modulation frequency along an acoustic frequency subband, or row, with approximate amplitudes.
- If modulation spectral features appear at more than one set of multiples along an acoustic frequency subband, then more than one source with a strong temporal pattern exists along that acoustic frequency subband.
- Modulation spectral features that appear "stacked" along a modulation frequency, or column, may belong to the same source if the modulation spectral features consistently appear at the same multiples.
- Sources with little-to-no temporal patterns will have modulation spectral features that appear "noisy" in the modulation spectrum, i.e. not appearing to have multiples along an acoustic frequency subband.

We note that the third property is conditional and is challenging in the case where sources overlap both in acoustic frequency subbands and in multiples of modulation frequencies. Also, grouping harmonics along a modulation frequency column at various acoustic frequencies sometimes may be an ambiguous process. We further discuss some of these concerns in our results in Section 2.3.3.

2.3.2 Unsupervised Identification Algorithm

Algorithm 1 summarizes the general implementation of our proposed method for finding sources based on the aforementioned properties of modulation spectral features. Two data structures $S(i, \vec{\omega}, m)$ and $V(j, \vec{\omega}, m)$ are used for storing information about each source found, where i references the source ID, $\vec{\omega}$ is a list of that source's acoustic frequencies (ω), m is that source's unique modulation frequency, and S and V are for sources with

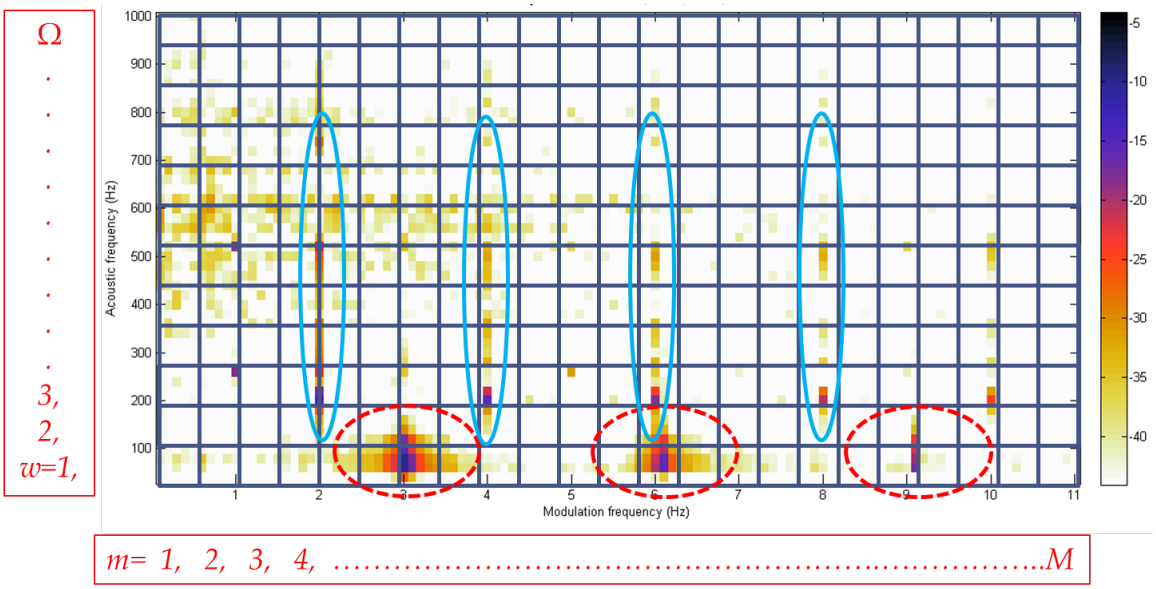


Figure 4: Modulation spectrum of a 5 second clip of a mixed music signal. Modulation frequencies appear at multiples for two sources with strong temporal patterns: snare drum (2 Hz) outlined by long, narrow ovals and bass drum (3 Hz) outlined by short, wider ovals. Also shown are "noisy" sources: two talking voices ranging from about 400 Hz - 900 Hz in acoustic frequency, which contain little-to-no temporal pattern. We demonstrate groupings of pixels to form blocks that may be used as bins.

strong and weak modulation, respectively. For our modulation amplitude peaks, we use thresholding to reduce less significant features in the modulation spectrum. This preprocessing/inhibition stage will promote large global peaks along subband rows and suppress weaker local peaks.

2.3.3 Experimental Results

Experiments for unsupervised source identification with our algorithm consisted of a variety of signals. We used 30 (24 synthetic, or manually composed, and 6 authentic, or professionally recorded) unique audio signals of 5-10 seconds in length with a sampling rate of 16 kHz. These audio signals were taken from commonly used datasets (MIREx, TIMIT, RWC, University of Iowa Musical Instrument Sounds Database, LabRosa, and TRIOS source separation dataset). The test signals consisted of sentences from multiple speakers, single musical instrument notes, harmonic song segments, percussive song segments, and mixtures. The modulation spectrum was computed (Modulation Toolbox for

Data: magnitude modulation spectrum $P(\omega, m, \alpha(\omega, m))$, $\forall\omega \in \Omega$, $\forall m \in M$
Result: S,V

```
begin
    initialize S with all sources S(i, ω, m) corresponding to the maximum amplitudes, max(α(ω, m)),
    along each row;
    begin at the initial bin, (ω = 1, m = 1);
    for ∀ω ∈ Ω do
        for ∀m ∈ M along row ω do
            if α(ω, m) ≤ threshold then
                ||α(ω, 2m) − α(ω, m)||_2 = A;
                ||α(ω, 3m) − α(ω, m)||_2 = B;
                if A ≃ B (amplitudes at second and third multiples of m are approximate) then
                    if ((ω − 1), m)∃S_i then
                        update existing source ID, S(i, ω⃗ + (ω − 1), m);
                    else
                        add a new source ID, S(i + 1, ω, m);
                    end
                else
                    add a new source ID, V(j + 1, ω, m);
                end
            end
        end
    end
end
```

Algorithm 1: Algorithm for unsupervised source identification using the modulation spectrum.

MATLAB) for each signal using rectangular windows, Hilbert demodulation, 20 Hz acoustic frequency sub-bandwidths, and modulation frequency resolution varying from 0.5 Hz to 1 Hz. As preprocessing for the algorithm, we used an amplitude threshold of the top 75 percent of components in the modulation spectrum for source identification. As mentioned in Section 2.3.2, this threshold inhibits less significant components, which typically have low amplitudes.

2.3.3.1 Verification Tests with Synthetic Signals

First, we verified the functionality of our algorithm with 24 synthetic signals we composed and pieced together ourselves in order to test signals with known acoustic and modulation content. The error rates and average accuracy for these tests are shown in Table 1. For unsupervised identification results of all 24 tests, we were able to achieve a high true-positive-to-ground-truth rate (where sources identified matched the pitch or harmonics of

Table 1: Total ground truth sources (GT) with true positive (TP), false positive (FP), and false negative (FN) sources identified (or not identified-FN) over 24 clips.

Overall (24 clips)	GT	TP	FP	FN
Total Sources	35	32 (91.4%)	31 (88.6%)	3 (8.6%)

the ground truth labelled sources) of 91.4 percent. Per each signal test, we individually calculated a true-positive-to-ground-truth rate and calculated an average of 97.6 percent. For our overall false positives (where sources identified did not match the ground truth sources), however, we noticed these sources often still consist of some of the acoustic harmonics from the ground truth sources. This observation indicates that the false positives were actually remnants from true positives that were not successfully grouped with true positives, especially in cases where sources had no harmonics (mostly percussive and noisy sources). We also note that these false positives may be due to sources that truly are present but are unable to be determined by an average human ear. Lastly, our algorithm shows a considerably low overall false negative-to-ground truth rate (where sources that are in ground truth were not identified at all) of only 8.6 percent, which is favorable. We attribute this small amount of error to some sources being quieter in volume and not meeting our preprocessing threshold filter, such as targeted sources that are simply overpowered by other sources. Also, since we are using 20 Hz acoustic frequency subbands, some sources may be grouped with other sources and are not being accounted for. We noticed that sources that overlap in both acoustic and modulation frequency may be separately identified if their acoustic harmonics are slightly different, which is attributed to the precision allotted by the resolution parameter. As for the noisy sources (speech, cymbals, etc.), these were either stored as noisy sources or identified as separated sources having various rhythms.

2.3.3.2 Resolution Tests with Authentic Signals

We extended our preliminary experiments to six sample clips from professionally recorded songs and varied the modulation frequency resolution from fine (0.5 Hz) to broad (1 Hz) to see how the unsupervised source identification changed. These songs (most of which are

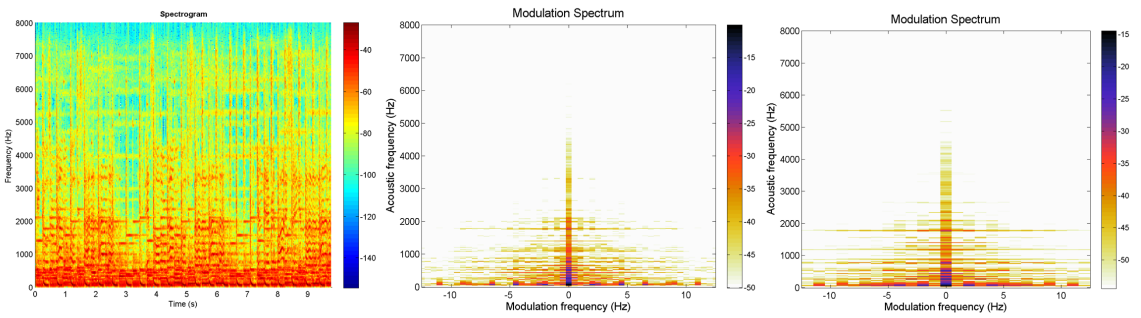


Figure 5: Traditional spectrogram, modulation spectra at finer (0.5 Hz) and broader (1 Hz) modulation resolutions (from left to right) of a clip from "Around the World" by ATC verified to contain bells, drums, and vocals.

generally well-recognized) were chosen due to their strong temporal characteristics and in an attempt to cover a variety of song styles/genres. They include "Around the World" by ATC (techno), "Sugar Plum Fairy" by Tchaikovsky (classical–the Romantic period), "The Corner" by Common (hip-hop), "Bad" and "Beat It" both by Michael Jackson (pop), and "Swing-style" by Willie Thompson (jazz–swing style). We (as trained musicians) manually labelled the ground truth sources for our signals beforehand. Figure 5 shows how the resolution of the modulation spectrum is affected by the fine and broad resolution settings (similar to [81]), thus allowing more or less room for identification precision. As shown in Table 2, the broad resolution results more closely match our ground truth sources. The broad resolution also eliminated the noisy sources that were identified with the fine resolution and identified most of them as being structured instead, which may be preferable in some cases. Since the noisy source identification (V) was not as useful, we may eliminate this portion of the algorithm in future work.

2.4 Conclusion

Modulation features have demonstrated promise and potential to automatically distinguish and identify information in music signals. The success of modulation features in other applications has motivated their use in music signal processing. Some of these applications include genre classification, emotion detection, source separation, timbre modeling, and

Table 2: Displayed in this table are the number of sources identified by the algorithm in the structured (S) category, meaning sources with strong temporal patterns, and in the noisy (V) category, meaning sources without strong temporal patterns. The numbers of identified sources are compared with the numbers of sources in the ground truth (GT), or hand-labeled, data for sample clips from six authentic (professionally recorded) signals of various genres at different modulation frequency resolutions ("Fine" at 0.5 Hz and "Broad" at 1 Hz). As shown, the "Fine" resolution results in more false positive identifications of structured sources while the "Broad" resolution results in a more accurate identification, despite identifying the ground truth, noisy sources as being structured as well.

	Fine		Broad		GT	
Song Clip (genre)	S	V	S	V	S	V
"Around the World" (techno)	8	1	3	0	1	1
"Sugar Plum Fairy" (classical)	1	0	2	0	1	1
"The Corner" (hip-hop)	6	1	4	0	3	1
"Bad" (moderate tempo pop)	11	0	4	0	2	0
"Beat It" (fast tempo pop)	10	0	3	0	3	0
"Swing-style" (jazz)	5	1	4	0	3	1

musical instrument recognition. The majority of algorithms and frameworks surveyed in this chapter either extract modulation features directly from the signal (as in the time domain or traditional frequency domain) or extract features from the modulation spectrum. The research reviewed in this chapter employed modulation features in ways similar to the way other common spectral features have been used. Advantages of modulation features include their long-term modeling capability, perceptual relevance, noise invariance, and visible characteristics.

Although modulation features offer a path towards invariant representations for music, they may suffer from underlying problems with sensitivity to parameter choices. These "parameters" are not merely numeric variables, such as filterbank bandwidths or modulation spectra resolution; they include implementation options such as the choice of envelope detection and demodulation methods. Perhaps these issues can be overcome by improving our fundamental understanding of modulation theory, straying away from simply reusing speech-tailored data mining techniques, incorporating adaptive learning methods, and directly targeting hardware implementations. Addressing these concerns will assist with furthering the intuitive understanding and development of modulation spectral features for

invariant representations of music. In the meantime, we demonstrate how to achieve optimal parameters and make choices depending on the signal type.

Lastly, we further exploit the usefulness of modulation spectral features by showing that the modulation spectrum facilitates the unsupervised identification of sources with little *a* priori information. This modulation spectral domain may be more useful for easily recognizing multiple sources with differing temporal patterns than the time domain and the traditional frequency domain. Even if the sources overlap significantly in acoustic frequency, the additional modulation frequency axis may signify that there are multiple sources present. Experiments confirmed capabilities of the unsupervised identification with signals containing multiple sources, whether percussive, harmonic, vocal, or mixtures. Future work may be to utilize our method as preprocessing steps for other music data mining tasks, such as unsupervised source separation via modulation filtering.

CHAPTER 3

EXPLORING FREQUENCY MODULATION (FM) FEATURES IN THE MODULATION SPECTRUM WITH APPLICATIONS IN VIBRATO ANALYSIS

[1]Modulation frequency features have shown promise in representing distinguishing characteristics of audio signals, and, in turn, have been used for audio data mining and classification algorithms. Previous research employing the modulation spectrum has focused on its representation of amplitude modulation (AM) characteristics in a signal, but little focus has been placed on the observation of frequency modulation (FM) structure in the modulation spectrum. This chapter focuses on developing an understanding of how simple signals, both AM and FM, are manifest in this modulation spectrum to provide deeper intuition and lead to wider application. In particular, we perform a case study on removing, synthesizing, and copying vibrato in trumpet notes to demonstrate the benefits of FM spectral analysis.

This chapter revisits general modulation theory (Section 3.1), demonstrates how FM features as well as AM features may be observed in the modulation spectrum, depending on resolution (Section 3.2), and discusses why such FM features may be useful (Section 3.3). Examples include simple, synthetic signals as well as more complicated music signals.

3.1 General Modulation Theory

We recall the general modulation theory from Section 2.1.2. Consider a simple modulated signal $x(t)$ with the general form $x(t) = m(t)c(t)$ for AM and $x(t) = c(t + Im(t))$ for FM, where I controls the amount of frequency variation. In some applications, such as AM or FM radio demodulation, the structure of $c(t)$ may be known, and the goal is to estimate $m(t)$. In other applications, $c(t)$ and $m(t)$ may need to be jointly estimated. In this more

[1]This chapter is modified from *Sephus, N., Lanterman, A., & Anderson, D. (2013). Exploring Frequency Modulation Features and Resolution in the Modulation Spectrum. In 2013 IEEE Digital Signal Processing (DSP) and Signal Processing Education (SPE) Meeting (pp. 169 - 174). Napa, CA.* [81] and an upcoming journal paper submission.

complicated case, we may seek $m(t)$ by detecting its slowly varying envelope, or we may first try to estimate the carrier signal $c(t)$.

In the simple cases described above, where the signal contains a single carrier modulated by a single modulator, the signal can be relatively easily decomposed. For more complex signals, such as music, multiple modulators and carriers are necessary, which are more difficult to estimate. Also, recall from Section 2.1.2 that the most well-known form of general decomposition involves filterbank analysis; the signal $x(t)$ is divided into k subbands $x_k(t) = x(t) * h_k(t)$, where the impulse responses of the bandpass filters are denoted as $h_k(t)$, and $*$ represents convolution. The resulting subband signals $x_k(t)$ are decomposed into modulators $m_k(t)$ and carriers $c_k(t)$ after undergoing envelope detection and carrier estimation [1].

3.1.1 Basic AM Example

Consider a simple, sinusoidal AM signal defined as $x(t) = \{A_c + \cos(\omega_m t)\}\cos(\omega_c t)$, where A_c is the amplitude of the carrier signal, $\omega_m = 2\pi f_m$, f_m is the modulator frequency (Hz), $\omega_c = 2\pi f_c$, and f_c is the carrier frequency (Hz). An example is shown in Fig. 1. As a nearly trivial example, consider a sum of two cosines. By a well-known trigonometric identity, this sum can be decomposed into an AM structure: $2\cos(\omega_c t)\cos(\omega_m t) = \cos([\omega_c - \omega_m]t) + \cos([\omega_c + \omega_m]t)$, with $\omega_c \gg \omega_m$. Our interpretation of these quantities depends on the application. In AM radio, we would think of a carrier of frequency ω_c being modulated by a signal of frequency ω_m. In the case of two musical tones of slightly differing frequency – such as when a musician tunes their instrument – we would perceive of $\cos(\omega_m t)$ as a "beat tone."

3.1.2 Basic FM Example

Recall the basic AM example from Section 3.1.1. Now consider a simple, sinusoidal FM signal $x(t) = \cos(\omega_m t + I\cos(\omega_c t))$ where $\omega_m = 2\pi f_m$, f_m is the modulator frequency (Hz), $\omega_c = 2\pi f_c$, and f_c is the carrier frequency (Hz). The instantaneous frequency is

$f_i = 2\pi f_c - I2\pi f_m \sin(\omega_m t)$. For higher modulator-to-carrier frequency ratios, FM spectral features may exhibit sidebands spaced at integer multiples of f_m on both sides of the carrier. The amplitudes of the resulting frequency components can be calculated via Bessel functions [82]. This chapter focuses on the relatively low modulation frequencies associated with vibrato, so we will not consider Bessel functions further.

3.2 AM/FM Features and Resolution in the Modulation Spectrum

Resolution of the modulation spectrum depends on its filterbank settings, such as the width of the subbands and whether or not the subband widths are uniform across acoustic frequencies. In addition to subband widths, three *windows* are typically considered when plotting the modulation spectrum. The first window is the length of the time-domain signal under analysis, whether it be the entire signal or a smaller segment. The second window is the length of the *acoustic* DSTFT. This second window is analogous to the previous window in that it may be thought of as the length of the segment of analysis in the signal's traditional spectrogram. An example of how subband widths may be small enough to show sum and difference tones or wide enough to merge these tones into one carrier tone is shown in Fig. 6. The final window is the length of the *modulation* DSTFT needed to compute the modulation spectrum, which can be compared to the N-length DFT used in a traditional spectrogram. These parameters determine the resolution, or "pixel size," of the modulation spectrum. Various interpretations of signals with similar modulation frequencies are demonstrated in Fig. 7. When plotting the modulation spectrum using the Modulation Toolbox in MATLAB [8], a pixel's vertical height across a given acoustic frequency row is equal to the width of the subband set for that acoustic frequency range. The pixel's horizontal width in the modulation spectrum is determined by a ratio of

$$2 \times \frac{subband\,width}{length\,of\,modulation\,DSTFT}. \tag{4}$$

The modulation DSTFT length may be greater than the subband width, which is sometimes referred to as decimation or downsampling. Other parameters may be incorporated, such

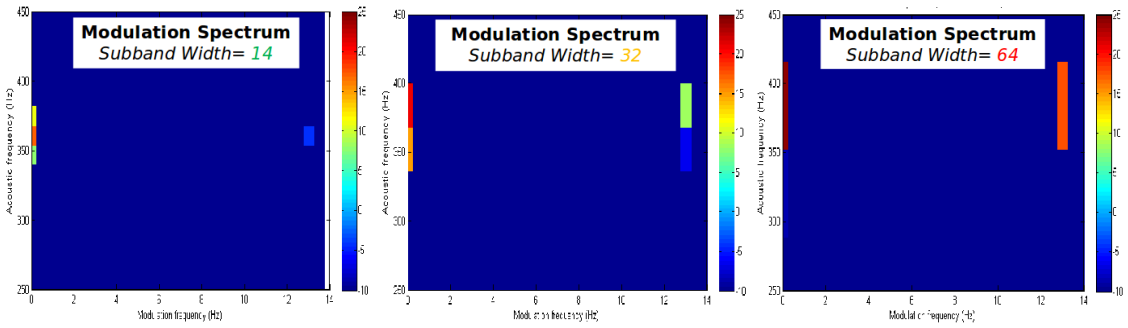


Figure 6: Changes in modulation spectra due to acoustic frequency subband widths (14 Hz, 32 Hz, and 64 Hz from left to right) for a carrier frequency of 370 Hz and a modulator frequency of 13 Hz. Sum and difference tones appear and gradually merge to one tone.

as window tapering and overlap amount. We must also choose a demodulation method. We use rectangular windows for our simple AM and FM feature analyses in the following sections.

3.2.1 Exploiting Hilbert Demodulation

We studied the Hilbert demodulation from [1] shown in Fig. 8, which results in real, non-negative envelopes. In contrast, "coherent demodulation" methods assume modulators are bandlimited with the possibility of being complex signals [77]. While coherent demodulation allows for better estimation of the carrier for nonstationary signals, the assumption of "coherently" bandlimited modulators is harsh for complicated signals, i.e. with many different overlapping modulators and carriers. Based on [1] and [77], Hilbert demodulation was not favorable for the frameworks geared towards speech applications. Since speech signals are nonstationary, they are better decomposed using a coherent demodulation method to better track the carrier (and, in turn, better track the modulator) as it changes over time. Despite the Hilbert demodulation method being unfavorable for most speech signals, we have found that Hilbert demodulation is indeed preferred for signals with strong temporal patterns, such as somewhat periodic segments of music, after observing their characteristics in modulation spectra. When experimenting with music signals, which are more complicated since they usually consist of multiple sound sources, we found it more difficult to "coherently demodulate" musical rhythms. In this case, we exploit the "incoherent" nature

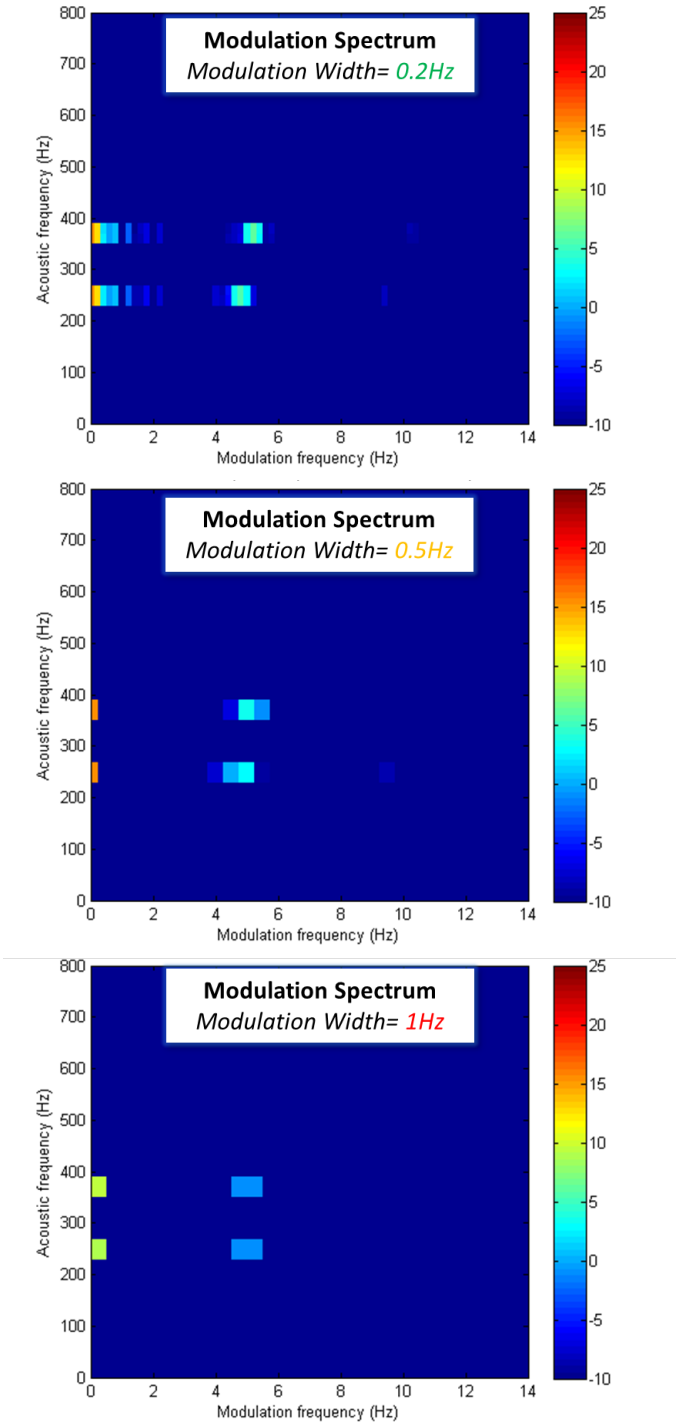


Figure 7: Changes in modulation spectra due to modulation frequency resolution (0.2 Hz, 0.5 Hz, and 1 Hz from top to bottom) for a signal containing two modulated carriers: one with carrier frequency of 370 Hz and a modulator frequency of 5.4 Hz and the other with carrier frequency of 250 Hz and modulator frequency of 4.8 Hz.

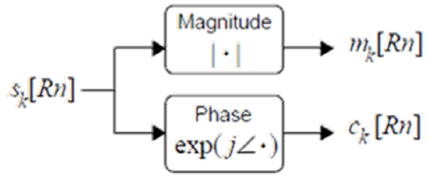


Figure 8: Hilbert demodulation method from [1]. $s_k[Rn]$ is the downsampled subband signal, $m_k[Rn]$ is the modulator signal, and $c_k[Rn]$ is the carrier signal.

of Hilbert demodulation and default to this more-applicable method.

Often times, a signal that is originally decomposed via a filterbank then followed by Hilbert demodulation may contain more unwanted artifacts in the reconstructed signal [77]. Hilbert demodulation may misinterpret negative envelope values as being positive values, as shown in Fig. 9 when assuming the excursion of the modulator is from 1 to -1. In this scenario, the modulation frequency is estimated to be twice the true value. This difference in interpreted modulation frequency, which is not emphasized in the literature, may be due to the value of amplitudes in the carrier signal or whether the original signal is detrended prior to analysis. If the model of the signal is $m(t)c(t)$, where $m(t)$ and $c(t)$ are simple sinusoids like that mentioned in Sections 3.1.1 and 3.1.2, an ideal envelope will appear to be vertically centered around the origin. However, if we model the signal as $[1+m(t)]c(t) = c(t)+m(t)c(t)$, we would have an ideal envelope above the origin. The latter would indicate no negative portions of the envelope, which would mean Hilbert demodulation would correctly interpret the true modulation frequency in the modulation spectrum (see example in Fig. 10).

For most natural signals, however, modulators are nonnegative. In more complicated signals, "harmonics", or multiples, of the true modulation frequency is present in modulation spectra. This misinterpretation is usually the cause of unwanted artifacts in the reconstructed signal. We note that the unwanted artifacts may be avoided by removing the additional multiples of the true modulation frequency, such as via filtering.

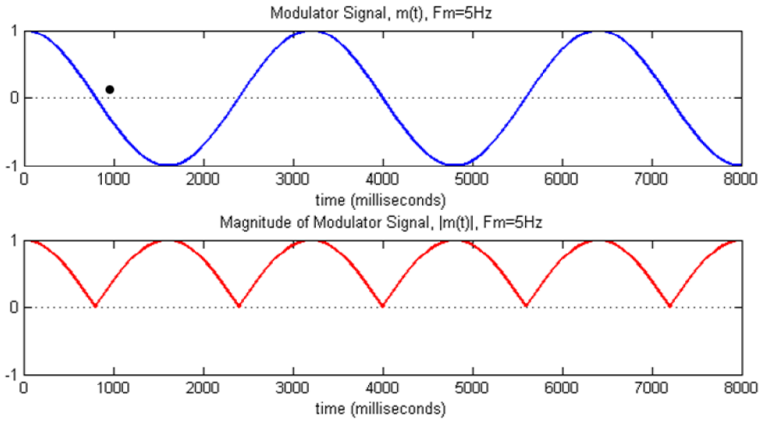


Figure 9: The true 5-Hz modulator (top) versus the modulator interpreted by Hilbert demodulation (bottom) with a frequency of 10 Hz.

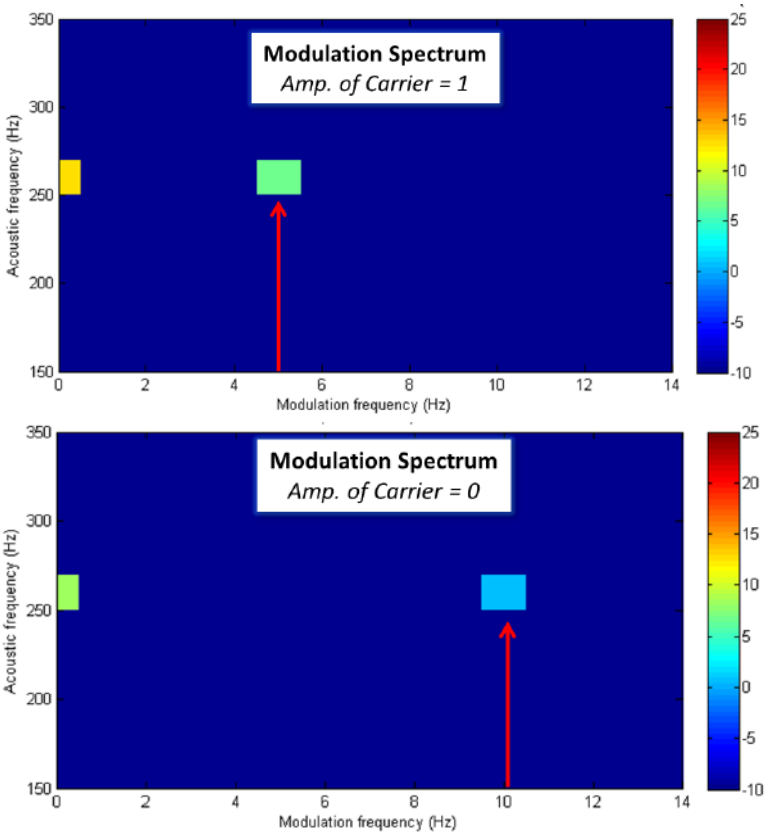


Figure 10: The true 5-Hz modulator (top) versus the modulator interpreted by Hilbert demodulation (bottom) with a frequency of 10 Hz.

3.2.2 AM Features in the Modulation Spectrum

As mentioned in Section 3.1.1, a sum of sinusoids may be perceived as either two distinct pitches (sum-and-difference tones) or as a beating tone (a single pitch with a discernible "tremolo" effect). These AM features can be well-observed in the modulation spectrum given appropriate resolution parameters for the filterbank and spectra. We consider fine, medium, and coarse resolutions. With fine resolution, the subband widths are small enough that both sum-and-difference tones and the carrier frequency are displayed in separate, non-overlapping pixels. As shown in Fig. 11, the four pixels that appear to have the highest dB energy are both sum-and-difference frequencies (383 Hz and 357 Hz located in the pixels above and below, respectively, the carrier pixel all along the acoustic frequency axis–0 Hz modulation) and the carrier frequency's modulation component (the pixel located at 13 Hz along the carrier frequency row). All portions are clearly shown since the center frequency of the sum-and-difference frequencies and carrier are non-overlapping with respect to the center frequency of the acoustic subbands. Otherwise, the modulation spectrum would show overlapping pixels. In Fig. 11, overlap occurs in the medium resolution example, where the carrier and its modulation component are grouped with the difference tone.[2] In the coarse resolution, subbands are wide enough to group all sum-and-difference tones, the carrier, and its modulation component into one longer pixel.

3.2.3 FM Features in the Modulation Spectrum

Typically, traditional spectrograms have been used to display FM sinusoidal signals. The modulation spectrum can be thought of as a DSTFT of the sinusoidal structure in an FM signal's traditional spectrogram when viewing the spectrogram analogously to the way we would view a time-domain plot of a sinusoid on an oscilloscope. Therefore, if resolution parameters are appropriately set, the modulation spectrum displays the rate of the repeated pattern(s) in the traditional spectrogram along each horizontal "slice," or subband row, of

[2] A similar example could demonstrate the carrier and its modulation component grouped with the sum tone as well. Either may occur, depending on the particular "bin" spacing.

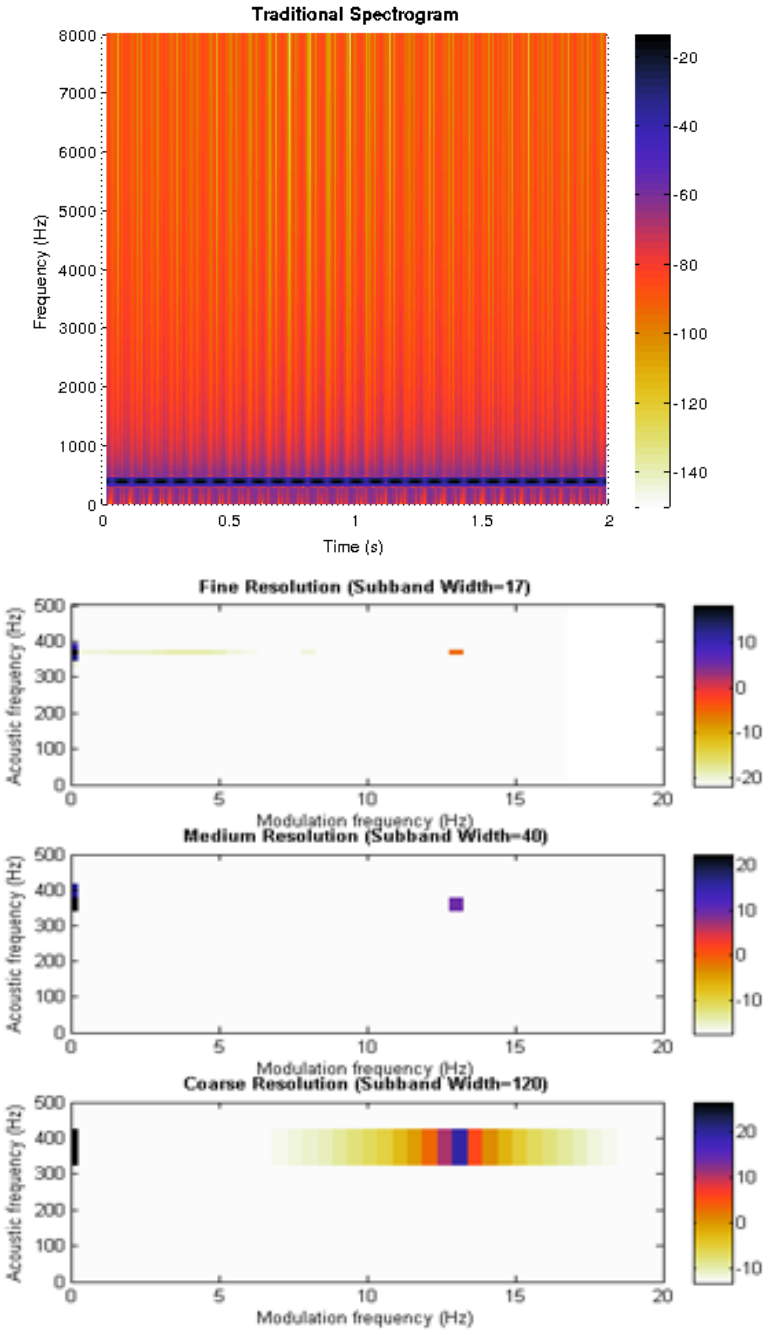


Figure 11: Example of a modulation spectrum of a simple synthetic AM signal (f_c = 370 Hz, f_m = 13 Hz, A_c = 1, and pixel width= 0.5 Hz) at different resolutions, i.e. subband widths.

acoustic frequency. For a sinusoidal-structured signal in the traditional spectrogram, the modulation spectrum would display the minimum and maximum acoustic frequencies described by the instantaneous frequency formula (previously described in Section 3.1.2) in "stacked" pixels at the signal's modulation frequency. Also, we would observe the modulation frequency of the median acoustic frequency at twice the original modulation frequency, since it occurs at twice the rate of the modulation frequencies along the maximum and minimum acoustic frequency subband rows. This FM signal analysis is shown in Fig. 12 with a more detailed view in Fig. 13. We immediately see that the fine resolution shown here would not be ideal for displaying these features. However, the medium and coarse resolutions display these features appropriately by showing the maximum and minimum acoustic frequencies from the traditional spectrogram (approximately 500 Hz and 200 Hz, respectively) at the original modulation frequency of 13 Hz. We also observe the median instantaneous frequency from the traditional spectrogram, which crosses the median acoustic frequency subband at approximately 350 Hz and occurs at twice the rate of the signal's original modulation frequency along the x-axis (26 Hz).

The resolution of the modulation spectrum may be adjusted to show more acoustic frequency excursions at other multiples of the original modulation frequency, as long as the subband widths in the modulation frequency are sized appropriately to group and display such detail. The FM signal example in Fig. 14 has a much larger I value, which causes more deviation of the instantaneous frequencies from the carrier frequency. In this example, the excursion of instantaneous frequencies "folds over" the zero-frequency line. In the traditional spectrogram, we can divide the subband rows into three distinct regions of frequencies according to changing rates of the repeated patterns: the lower region from 0-500 Hz, the middle region from 500-1000 Hz, and the upper region from 1000-1800 Hz. The corresponding modulation spectrum shows high-intensity modulation frequency at approximately 26 Hz for the lower region, 13 Hz for the middle region, 39 Hz near the peak of 1000 Hz, and 13 Hz for the upper region.

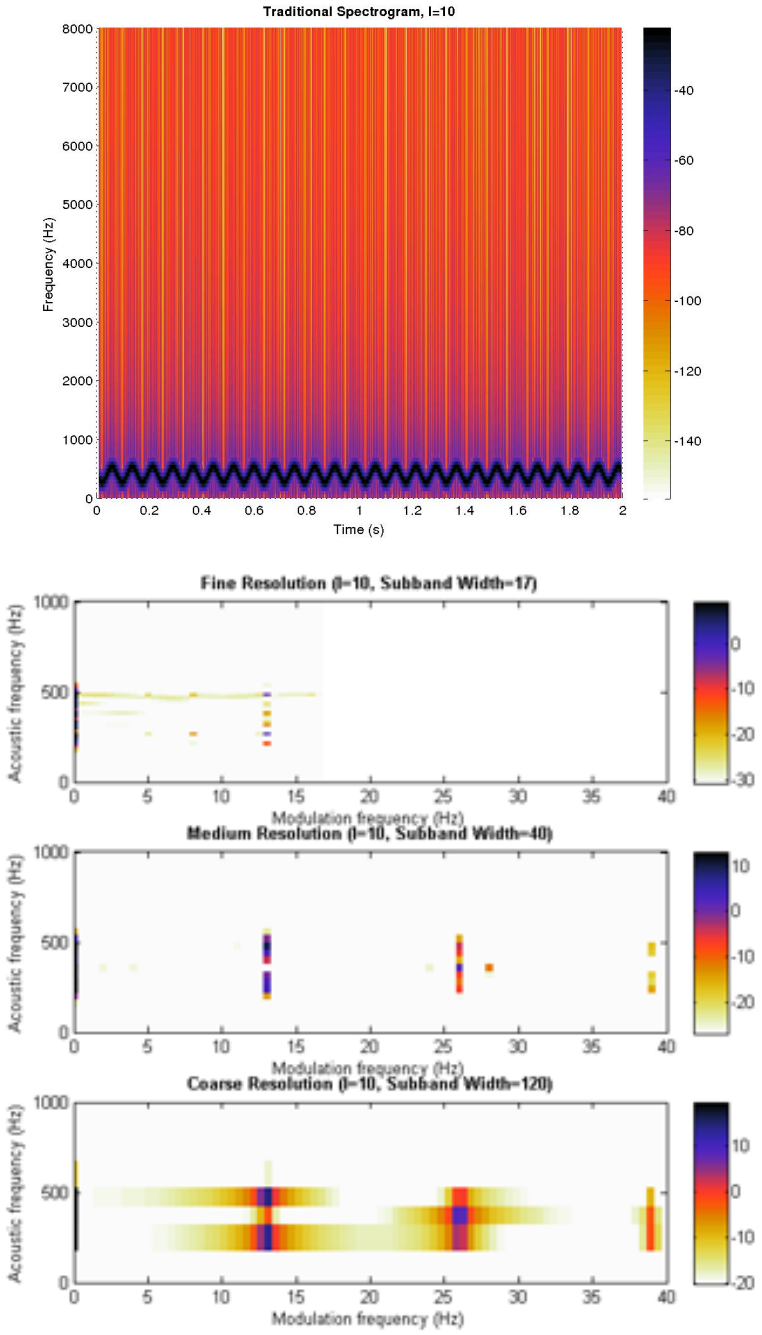


Figure 12: Example of a traditional spectrogram of a simple synthetic FM signal (f_c = 370 Hz, f_m = 13 Hz, I = 10, and pixel width= 0.5 Hz) and its corresponding modulation spectrum at different resolutions, i.e., various subband widths.

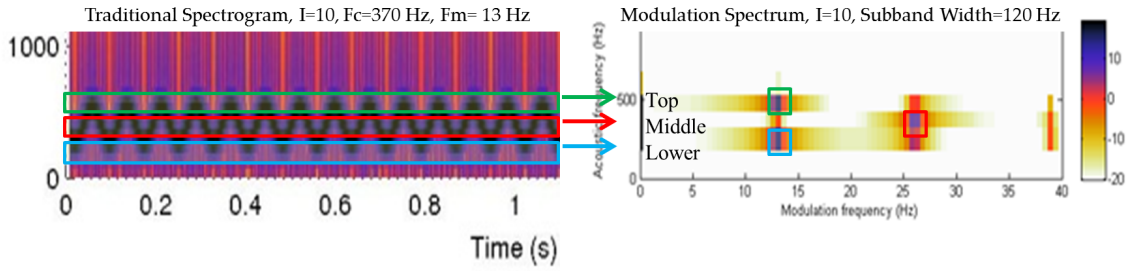


Figure 13: A traditional spectrogram and its corresponding modulation spectrum (from the signal used in Fig. 12) demonstrating the top, middle, and lower frequency subbands at a coarse resolution, i.e., wider subband widths.

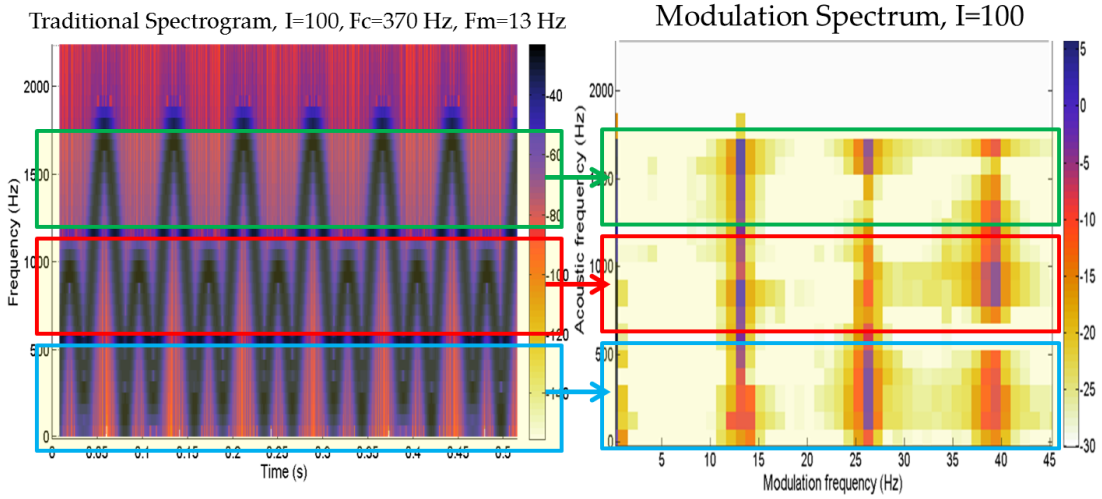


Figure 14: Example of a traditional spectrogram of a simple synthetic FM signal (f_c = 370 Hz, f_m = 13 Hz, I = 100, and pixel width= 1 Hz) and its corresponding modulation spectrum at a coarse resolution.

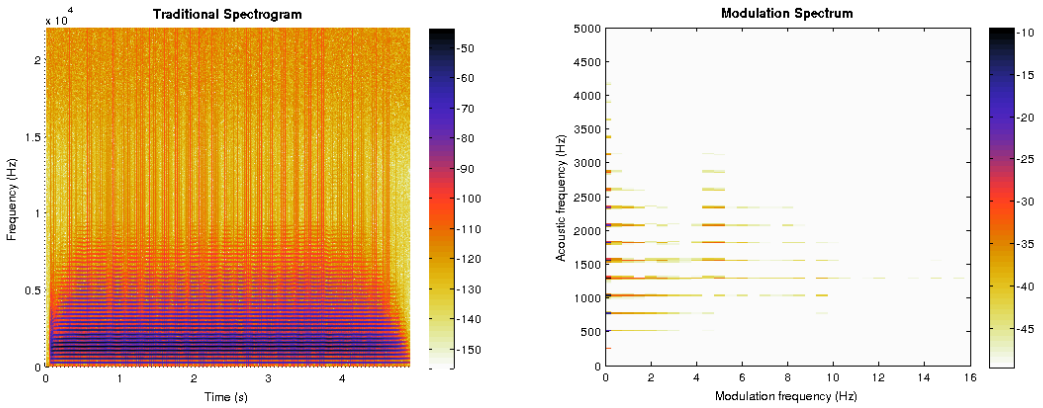


Figure 15: Example of a traditional spectrogram (left) containing a 2-second, vibrato-style trumpet note at middle C (pitch frequency of approximately 262 Hz) with an approximate 5 Hz vibrato at each of its harmonics and its corresponding modulation spectrum (right).

3.3 Applications of FM Features in the Modulation Spectrum

3.3.1 FM Features in Music

The advantages of exploiting the modulation spectrum for modulation spectral features may be useful in music information retrieval. Other forms of modulation features have been previously used for robust genre classification [58], emotion detection [63], timbre modeling, and musical instrument identification [52].

Two examples of FM features in the modulation spectrum are demonstrated in Figs. 15 and 16. The first example demonstrates how the FM content of a trumpet note[3] with approximately 5 Hz vibrato is manifest in the modulation spectrum.

A slightly more complicated example demonstrates FM features where resolution plays an important role. Fig. 16 shows a 2-second recording of a chord played on a Hammond organ, which is known for its richly modulated sound. Modulation components in the lower frequencies are barely visible in the traditional spectrogram, and the amount of modulation is difficult to determine. The modulation spectrum indicates more clearly the modulation along each acoustic frequency subband.

[3]This trumpet note was taken from the University of Iowa Electronic Music Studios' Musical Instrument Samples found at http://theremin.music.uiowa.edu/MIS.html.

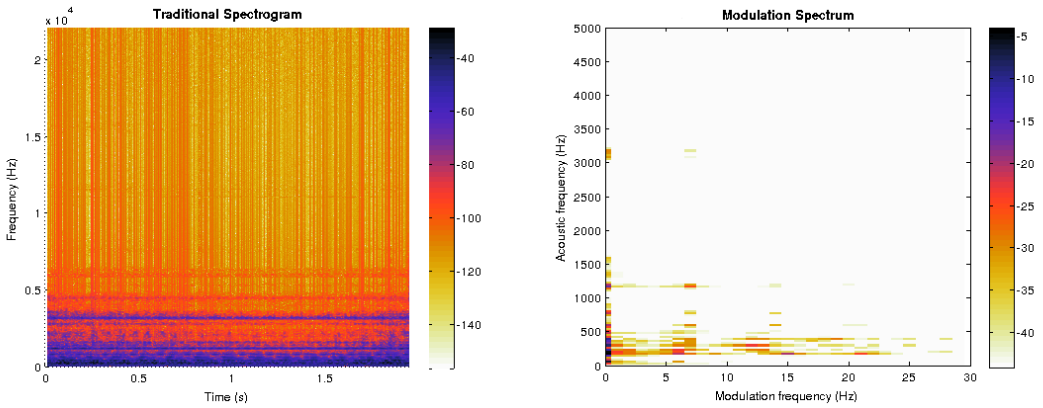


Figure 16: Example of a traditional spectrogram (left) containing a 2-second chord played on a Hammond organ and its corresponding modulation spectrum (right).

3.3.2 Vibrato Modification/Synthesis

We now apply this analysis to vibrato synthesis and modification in music signals as an artistic tool, illustrated via trumpet sounds. We start with two signals of single, same-pitch trumpet notes, each about 2 seconds long, sampled at 16 kHz. The first signal has little-to-no vibrato, and the second has much more vibrato. We can decompose, demodulate, and plot a modulation spectrum for each. As shown in Fig. 17, along the acoustic frequency subband rows, frequency modulation is more apparent in the signal with more vibrato, i.e more modulation components are visible. We use these two signals to demonstrate how to remove, synthesize, and copy vibrato.

We first show how to remove vibrato from the second signal, which has more vibrato, as shown in Fig. 18. After decomposition via a filterbank with 20-Hz subband widths, we identify which modulation components we want to remove. In this case, we choose the 5 Hz modulation component from the signal at the 260 Hz acoustic frequency subband since it is identified as the highest amplitude. We then select a subset of modulators from other subbands that best match in harmonics. We could either filter the modulators along subbands to *stop* frequencies around 5 Hz or filter the modulators on subbands to *pass* frequencies below 1 Hz. The second option ideally would remove all modulation components and leave only the carriers. We choose this option and follow with reconstruction of the

signal by multiplying modulators with carriers and summing all subband signals in the time domain. Attempting to remove, or filter, vibrato from a portion of a signal that contains no vibrato (i.e. attempting to remove a modulation component that does not exist) can cause "choppy"-sounding distortion in the reconstructed signal.

Next, we illustrate how to synthesize vibrato in the first signal, which has hardly any vibrato, as shown in Fig. 19. We start by decomposing the signal via a filterbank with 20-Hz subbands and demodulating the signal into modulator and carrier pairs on subbands. We assume that the modulators in signals without much vibrato are mostly flat. We then multiply a synthesized cosine signal by the carrier signal located on the 260 Hz subband to give it the characteristics of an envelope with some preferred frequency and amplitude. Note that the modulation frequency range for "tremolo" is defined as being 4-8 Hz and "roughness" is defined as being 8-10 Hz [5]. For this example, we use a 5 Hz cosine signal, similar to what we found in the vibrato signal during the previous example of removing vibrato. We want to synthesize this at the 260 Hz acoustic frequency subband, but we also want to perform this operation along a subset of other modulator signals that appear to be similar, such as the harmonics. After reconstructing the signal, the resulting signal sounds as if vibrato is present. In this case, the resulting signal sounds "too perfect," as in not sounding very natural.

Lastly, we supposed it might be better to provide a more natural-sounding vibrato by copying modulation characteristics from a skilled human player employing vibrato onto the performance without vibrato. This example is demonstrated in Fig. 20. We first identify the highest modulator frequency (or frequencies) in the vibrato model signal. We then identify a subset of the modulators in the minimal-vibrato sample that are similar to those at its replacement-frequency modulators, and we replace this subset of modulator signals in the minimum-vibrato sample with those from the notable-vibrato sample. After reconstruction, this version may sound more natural than the vibrato synthesis method mentioned early in this section.

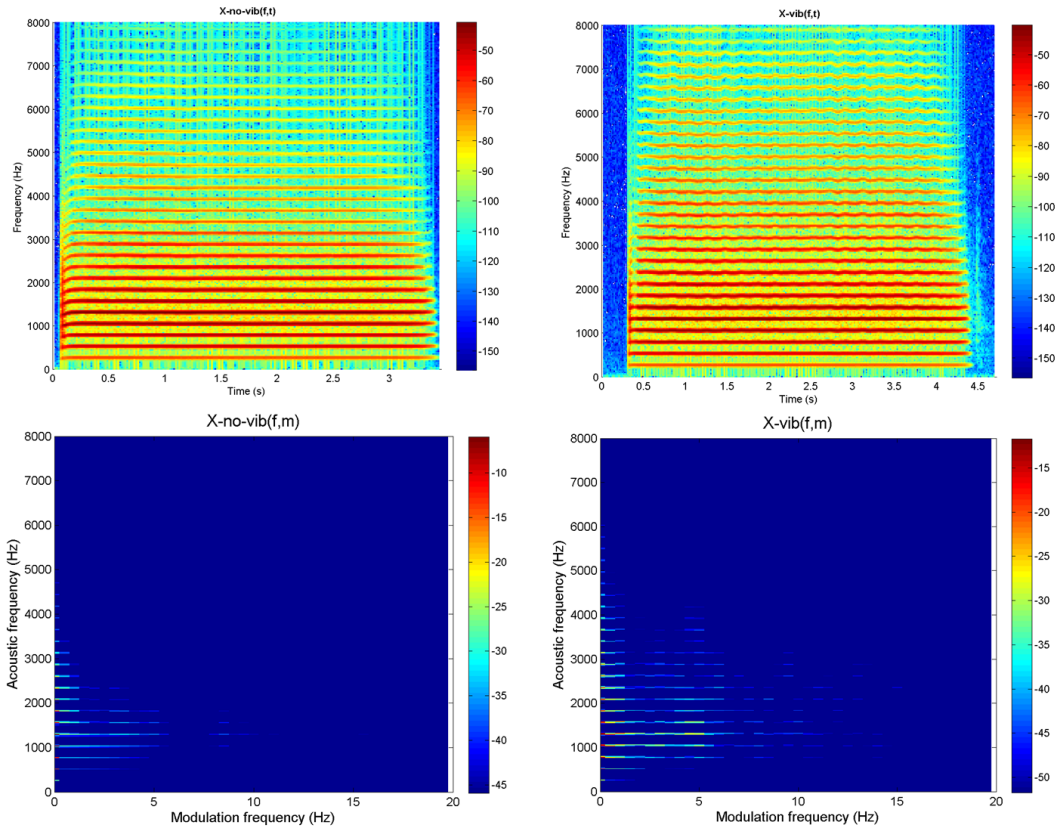


Figure 17: Traditional spectrograms (top) and modulation spectra (bottom) of two single trumpet notes, one with little-to-no vibrato (left side) and one with much more vibrato (right side).

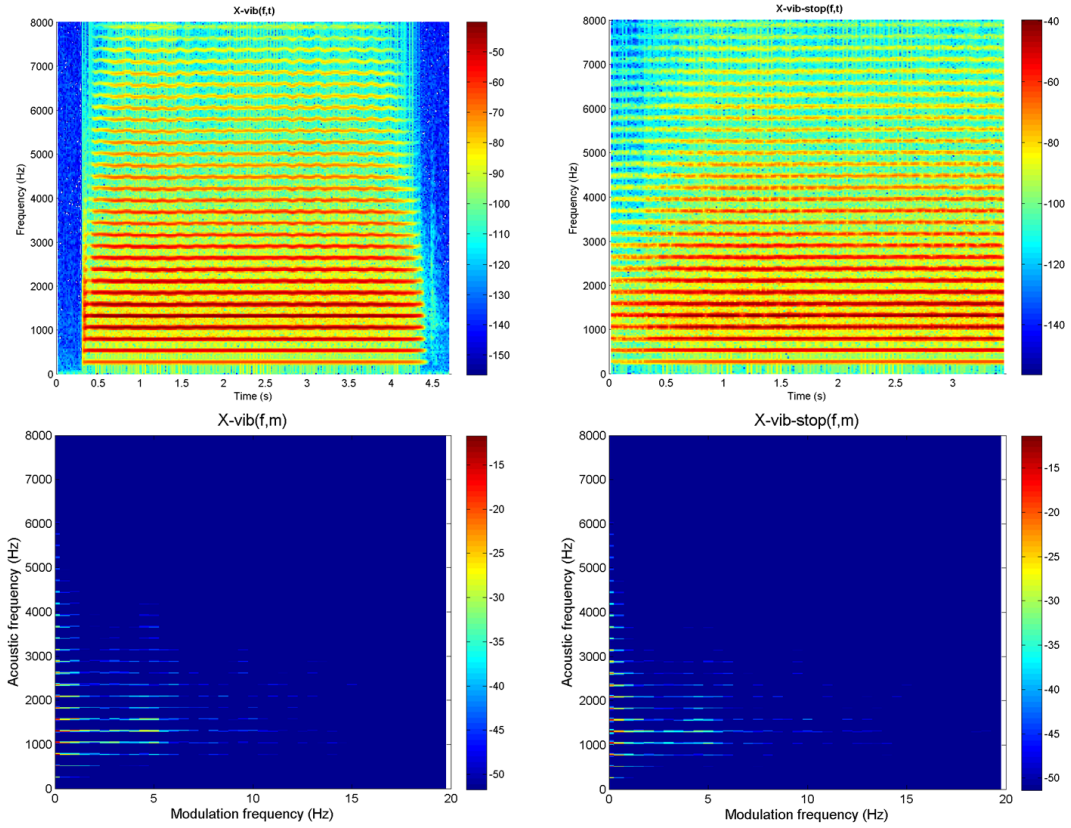


Figure 18: Traditional spectrograms (top) and modulation spectra (bottom) of two single trumpet notes, the first with vibrato (left side) and the second with an attempt to remove vibrato from the first signal (right side).

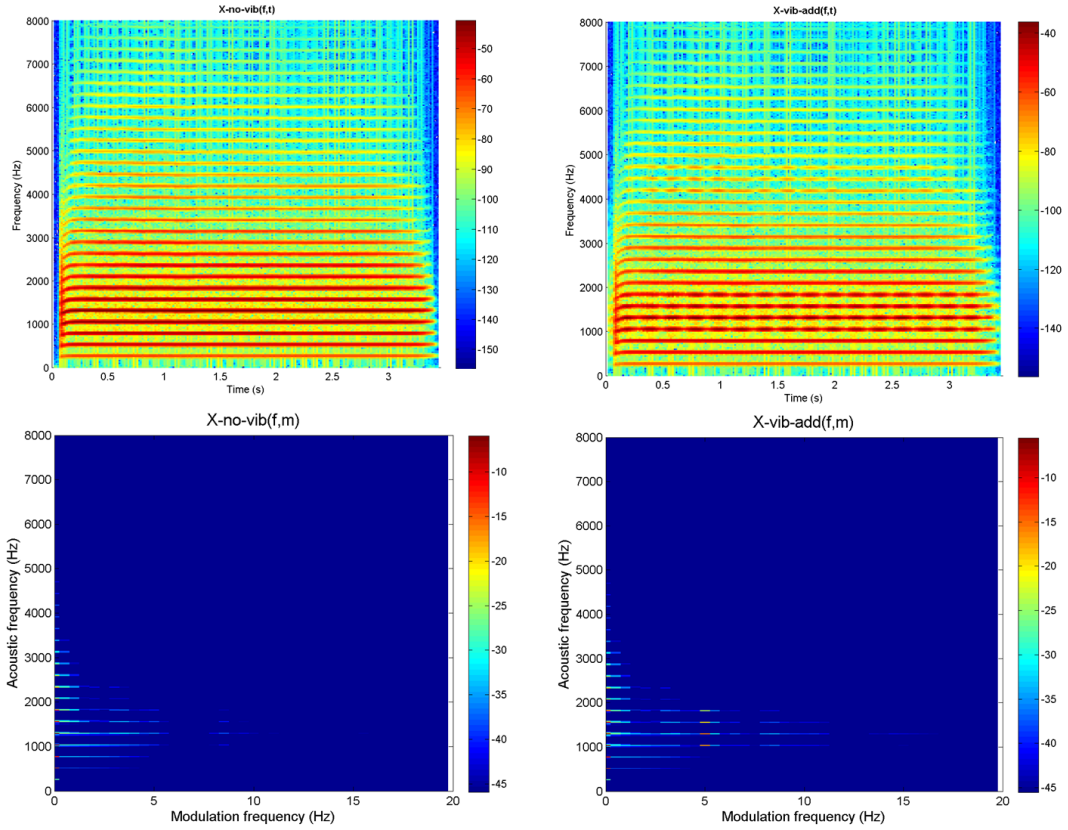


Figure 19: Traditional spectrograms (top) and modulation spectra (bottom) of two single trumpet notes, the first with little-to-no vibrato (left side) and the second with an attempt to synthesize vibrato in the first signal (right side).

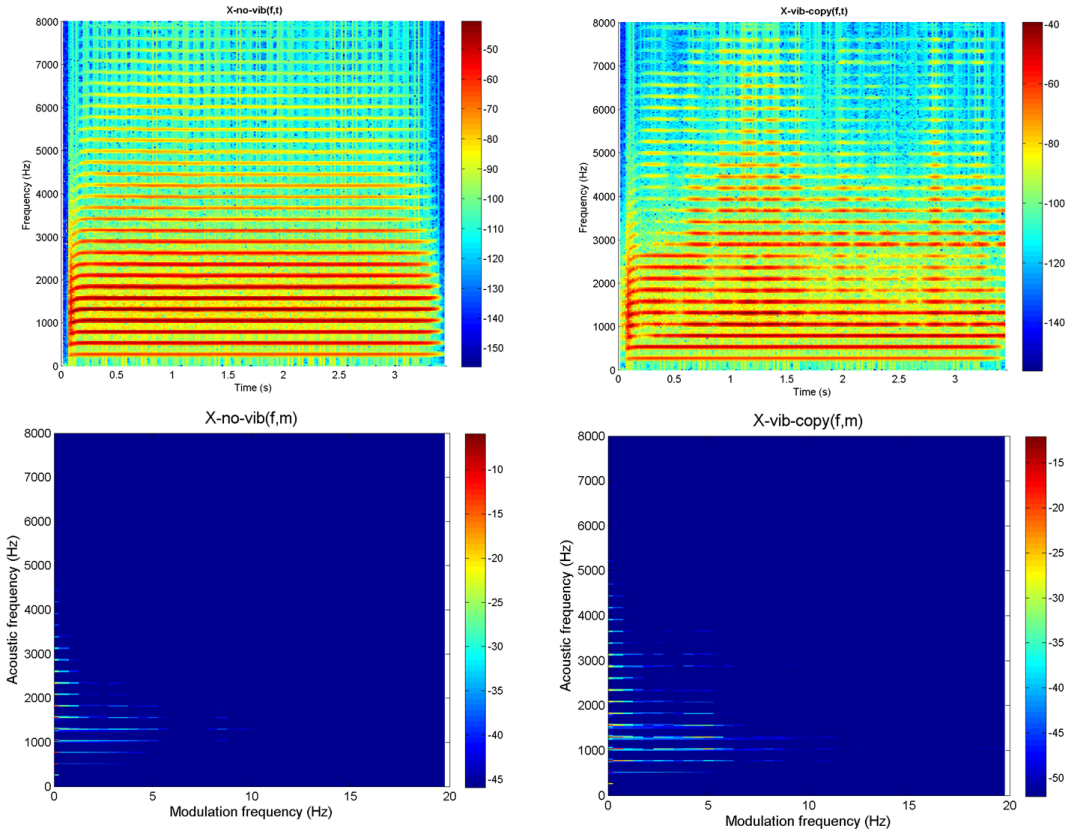


Figure 20: Traditional spectrograms (top) and modulation spectra (bottom) of two single trumpet notes, the first with little-to-no vibrato (left side) and the second with an attempt to copy vibrato onto the first signal (right side).

3.4 Conclusion

By gaining a deeper understanding of how the modulation spectrum can be a supportive visualization analysis tool for AM/FM signals, research using this framework may lead to more intuitive applications. Although the modulation spectrum is naturally well-suited for displaying and extracting AM features, it can be a useful tool for revealing FM features as well. We have demonstrated how a modulation filterbank framework may be adapted to analyze, remove, synthesize, and copy vibrato. Future work employing these spectral features may involve audio data mining, improved modulation filtering, as well as applications in other types of signal processing that utilize similar modulation characteristics. Also, with appropriate parameters for the modulation spectrum, a perceptual mapping for AM/FM signals may be explored.

CHAPTER 4

FRAMEWORK METHODOLOGY WITH APPLICATION TO UNSUPERVISED SOURCE SEPARATION IN MUSIC

[1]Modulation spectral features reveal information about the lower frequencies, i.e. "modulation frequencies," of temporal patterns that may occur across larger frequency subbands in a signal and may be visually observed via the modulation spectrum. We have developed a framework for exploiting these features in unsupervised source separation over song segments in music, where sources have strong temporal patterns. Our framework originates from, but contrasts with, a previous modulation filterbank framework for speech. We also provide case studies from publicly-available datasets and analyze results from listening tests. For a variety of monaural, polyphonic music signals, our framework allows for easy extraction of modulation spectral features for short-time and long-term analysis, which may be useful for unsupervised source separation and other preprocessing tasks in music.

This chapter is divided into the following sections. Section 4.1 discusses origins for the framework, motivations, and related work. Section 4.2 explains the methodology of the theoretical framework, and Section 4.3 analyzes various case studies and results of our listening tests. Section 4.4 summarizes our framework's capabilities, suggests future work, and provides a link to example sound clips.

4.1 Background

4.1.1 Exploiting Modulation Features

After finding the modulation spectrum of a signal, modulation spectral features, which include the modulation components' modulation frequency, acoustic/carrier frequency, and intensity, may be extracted to quantitatively summarize interesting aspects of the signal. Modulation spectral features have particular advantages, including perceptual aspects due

[1]This chapter is modified from an upcoming journal paper submission.

to their close relationship with the human auditory system [18], ability to represent a signal's long-term rhythmic structure and temporal/spectral patterns (as opposed to short-term features, such as MFCCs), noise invariance when identifying modulation characteristics [30], and ability to more directly visualize a signal's modulation characteristics via the modulation spectrum.

Section 2.1.6.3 considered existing methods that employed modulation features for source separation. We conclude that relatively few modulation feature frameworks focus on music features. Out of those frameworks that do focus on music, they are generally limited by their need for prior information. We provide explanation and motivation behind the development of our framework.

4.1.2 Motivation

Our framework is partly motivated by a similar method called harmonic-percussive sound separation (HPSS) [11]. The HPSS method consists of an algorithm designed to identify and later separate harmonic and percussive sounds by scanning the traditional frequency-versus-time spectrogram for visual differences between the two sounds, with harmonic components being horizontally smooth and percussive components being vertically smooth. This method involves iteratively minimizing an objective function that models this visual anisotropy as cost functions for the two types of components. Another similar technique quantifies the rhythmic similarity of music by a method involving dynamic periodicity warping [79].

We extend the concept of *longer-term* analysis windows so that temporal patterns are visible in the modulation spectrum, thus allowing us to automatically recognize the presence of certain sources in music signals. We employ this information to separate a signal or to selectively filter out or pass certain modulation components in a signal. This

work is related to that of nonnegative matrix factorization (NMF) and statistical modeling [83], [84], [85], and the algorithms implemented by participants of the community-based challenge named Signal Separation Evaluation Campaign (SiSEC)[2] (although these algorithms are usually targeted for particular separation tasks and assume some information about the sources is known). In particular, our separation framework is most related to that in [85], where the authors perform unsupervised source separation via estimation of multiple fundamental frequencies. This separation technique is based on nonparametric Bayesian extensions of NMF called infinite composite autoregressive models (iCARMs).

In contrast to the original modulation filterbank framework [1] and the accompanying masking technique in the modulation spectral domain, we focus more on grouping, or clustering, similar modulators along subbands directly in the time domain after automatically identifying sound sources from the modulation spectral domain, using this information as a guide in reconstructing, or synthesizing, resulting signals. Filtering and inverse transformation from the modulation spectral domain back to the time domain can result in artifacts that affect the quality of the reconstructed signal. In addition, some methods of the framework may be acceptable in speech applications but unacceptable in music applications. Our framework attempts to overcome some of these issues to be suitable for performing data mining tasks in music. Given a set of default parameters and a signal with sound sources containing temporal patterns, our framework is capable of performing unsupervised source separation in sound sources.

4.2 Framework Methodology

Our framework consists of three stages, as shown in Fig. 21: decomposition, identification, and reconstruction. The decomposition stage uses a filterbank to decompose a signal that is assumed to be either periodic or somewhat periodic (meaning there exists at least one sound source with a strong temporal pattern). The signal is filtered into acoustic, or carrier,

[2] http://sisec.wiki.irisa.fr/tiki-index.php

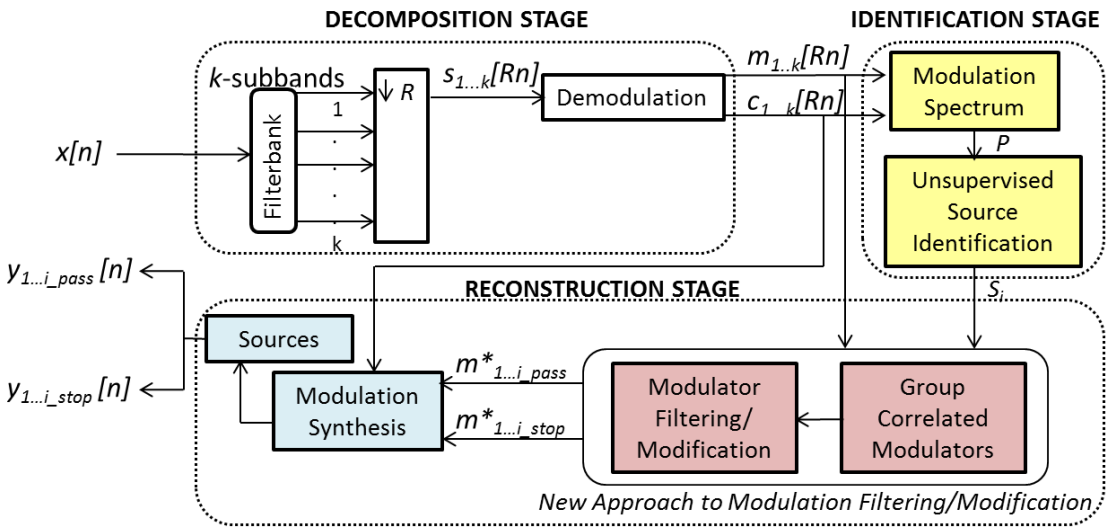


Figure 21: Framework for exploiting modulation spectral features in music data mining.

frequency subbands. The subband signals are then downsampled, to conserve memory and reduce computation time, and demodulated. By demodulating the subband signal into a modulator, or slow-varying envelope, and a carrier signal, we are able to plot a modulation spectrum, which we use in the identification stage. Based on the parameter study in [81], default parameters are chosen to establish the modulation spectrum resolution that may illustrate significant aspects of music signals. This is discussed in detail in Section 4.2.4. Previously mentioned in Section 2.3.1.1, the identification stage recognizes sound sources with temporal patterns by exploiting properties we know about observing music in the modulation spectra [3]. Lastly, the reconstruction stage utilizes the information from the identification stage to separate or modify the signal. This stage uses a correlation between modulator signals in the time-domain and/or filtering of modulator signals to reconstruct the signal of either a separated target source or synthesized source. We discuss the stages and parameters in more detail in the remaining subsections.

4.2.1 Decomposition Stage

Decomposition of a signal involves filterbank analysis (as mentioned in Section 4.1), demodulation, and representing a signal in its modulation spectrum. The modulation spectral

domain may visibly reveal information important for our task more directly than a traditional frequency-versus-time spectrogram. Currently, we use the Hilbert demodulation method from [1] described in Fig. 8. This type of demodulation is well-suited for the musical data mining tasks we explore. Most music contains sources that repeat periodically, so the assumption that modulators in music signals are real and nonnegative is appropriate. Also, this type of demodulation is simple and straightforward to implement, and has faster processing time, compared with other methods [80]. The properties of modulation spectral features, as demonstrated in Fig. 4 and restated from [3], are as follows:

- Modulation spectral features of a source with strong temporal patterns will appear at multiples of its modulation frequency along an acoustic frequency subband, or row, with approximate amplitudes.

- If modulation spectral features appear at more than one set of multiples along an acoustic frequency subband, then more than one source with a strong temporal pattern exists along that acoustic frequency subband.

- Modulation spectral features that appear "stacked" along a modulation frequency, or column, may belong to the same source if the modulation spectral features consistently appear at the same multiples.

- Sources with little-to-no temporal patterns will have modulation spectral features that appear "noisy" in the modulation spectrum, i.e. not appear to have multiples along an acoustic frequency subband.

4.2.2 Unsupervised Identification Stage

The identification stage involves extracting modulation feature information from the modulation spectra, i.e. finding where the dominant modulation frequencies are present along each row of the modulation spectrum and grouping harmonics that appear along rows and columns. The proposed algorithm from Section 2.3.2 embodies the general implementation

of the unsupervised identification stage using properties from the previous section. Preliminary experimentation presented in Section 2.3.3 demonstrates that no prior information such as number of sources or source pitches in a music signal is needed. We incorporate this algorithm in the identification stage of our overall framework. Unlike the previous version of the algorithm, however, the current algorithm no longer attempts to store information about sources without temporal patterns (since this information was sparse and not very applicable for our framework). For each source identified with strong temporal patterns, the current algorithm collects information that includes a list of that source's acoustic frequencies and unique modulation frequency. Nonetheless, we do continue to use thresholding to reduce less significant features in the modulation spectrum (discussed more in Section 4.2.4.4). This preprocessing/inhibition will promote large global peaks along subband rows and suppress weaker local peaks.

4.2.3 Reconstruction Stage

By knowing the acoustic frequency and modulation frequency information for each source found in the identification stage, we know which time-domain modulator signals to use for signal reconstruction. To achieve better-perceived quality in separated sources, the decomposition stage may be revisited after the identification stage and prior to reconstruction. By decomposing the signal into slightly broader subbands than what was used in the initial decomposition stage for identification purposes (discussed later in Section 4.2.4.3), less of the original source signal is decomposed. In turn, more of the reconstructed source signal is in tact. Similarily, by decomposing the signal into slightly smaller subbands than what was used in the initial decomposition stage for identification, we may overcome the issue of separating sources that overlap in acoustic subbands.

The reconstruction stage varies slightly depending on the task being performed. For separation, a pair of signals are produced for each source found (via masking of non-target frequency subbands) in the identification stage: the "pass" signal, which attempts to isolate a target source, and the "stop" signal, which consists of the remaining sources after

attempting to block the target source. In some cases, the sound sources may sound as if they are divided amongst the "pass" and the "stop" signal pair, in which the parameters would need to be modified to achieve desired results (discussed in Section 4.2.4).

For other tasks, such as removing or enhancing modulation in a signal, the signal's targeted modulation components are determined in the identification stage. These components may be removed by filtering the non-target frequencies out of the time-domain modulator signal via a bandstop filter. Before reintroducing the modulator to its carrier, we could perform some modification on the modulator such as filtering, masking in frequency, attenuating, amplifying, synthesize, etc.

As illustrated in Fig. 22, similar modulator signals suggest similar source origins. By finding the Pearson's correlation coefficient between each modulator signal, only the modulator signals that are most correlated with the target sound are "grouped" and used in reconstruction for that target sound. This, in turn, may overcome issues with the identification stage missing some useful subband signals needed for better sound quality in the reconstructed, separated signals (more discussed in Section 4.2.4.5).

4.2.4 Parameters and Settings

Defining the set of default parameters is important since the framework allows for unsupervised analysis. Knowing the purposes of these parameters may lead to greater intuition about how they may be adjusted from these default values to refine results. The parameters are shown in the diagram in Fig. 23, where the first and fourth parameters affect decomposition, the second and third parameters affect identification, and the last parameter affects reconstruction.

4.2.4.1 Parameter 1: Acoustic Frequency Subband Widths for Modulation Spectra

As described in Section 4.2.1, the initial decomposition stage prior to the identification stage involves constructing a signal's modulation spectrum. The acoustic subband width

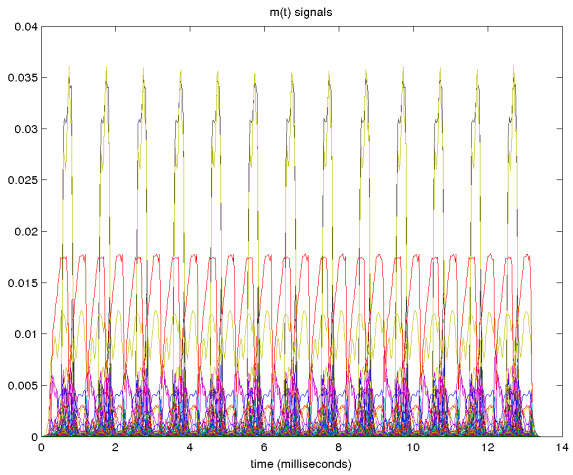


Figure 22: Modulators from every acoustic frequency subband are plotted in the time domain. This 8-second sample clip contains two sources, one percussive and one harmonic, that overlap temporally by repeating every half second. The general outline, or shape, of modulators belonging to a particular source look similar although the amplitudes may be different. Grouping, or clustering, modulators with similar shapes is done by calculating the correlation coefficients between our target modulator(s) and all others. Only those modulators with the highest correlation, based on a threshold, are used in reconstruction.

parameter affects the resolution along the y-axis in the modulation spectrum, thereby affecting which sources appear in the modulation spectrum.

Ideally, the acoustic subband widths of the acoustic frequency subbands must be wide enough to contain the full extent of an envelope signal, or modulator, from a particular sound source that exists along those acoustic frequencies. For example, a harmonic instrument may have four harmonics along a number of frequency subbands, in which an ideal acoustic subband width for a filterbank would isolate each harmonic into a separate subband. In the presence of multiple sound sources, an ideal acoustic subband width would isolate each harmonic or percussive sound source into separate subbands. Figure 24 shows how the subband widths affect the visual observation of modulation features in modulation spectra.

If prior information is known about a given signal, such as the pitch content of each present sound source within the analysis window, the acoustic subband widths may be set

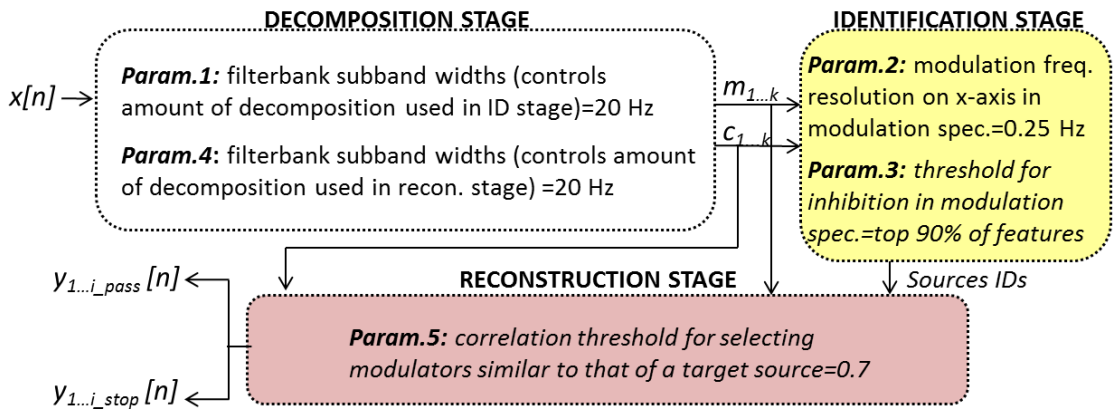


Figure 23: Description of framework parameters, as well as their default values, numbered according to the order in which they are used.

non-uniformly to better accommodate the frequency content. However, this type of information is not usually readily available, even in most publicly available datasets. Without this prior information, we found the best method for an unsupervised framework was to use uniformly-spaced acoustic subband widths. From a practical standpoint, less importance is placed on isolating harmonics from the same sound source, while more emphasis is placed on ungrouping different sound sources that occur in the same acoustic frequency subband.

We can consider how parameters may be adjusted if our default value does not reveal sufficient discriminating information in the modulation spectrum. If the reconstruction of a target source sounds as if parts are missing, such as notes missing from the melody, then the acoustic subband width may be widened to include the neighboring frequencies included in the missing parts. In contrast, if the reconstructed target signal sounds as if extra sounds are present, then the acoustic frequency subband may be made more narrow to ungroup the unwanted, non-target sounds. Our default value for this parameter is 20 Hz, which was determined to be a good starting point after experimentation with a wide variety of polyphonic, monaural music.

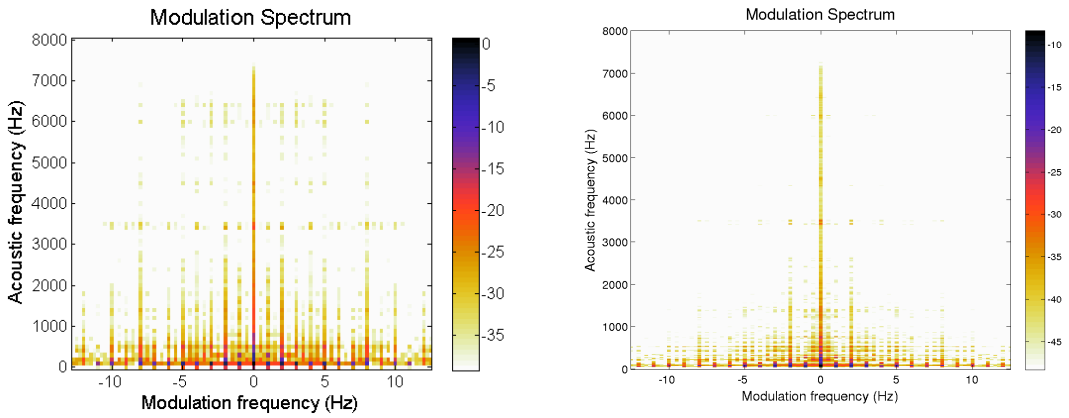


Figure 24: Modulation spectra of the same music signal with acoustic-frequency sub-bandwidths of 100 Hz (left) and 20 Hz (right).

4.2.4.2 Parameter 2: Modulation Frequency Resolution for Modulation Spectra

Section 4.2.2 discussed how our framework uses modulation spectra to identify sounds with strong temporal patterns. In Section 2.2, we discussed how the modulation frequency resolution determines the precision of the measurements of modulation frequencies. This resolution may be increased to cluster patterns into one pixel or decreased to show more detailed modulation content, as shown in Fig. 25. In some cases, coarse resolution may be favorable for reducing the effect of less significant components, or those with slight misalignments in tempo, in the identification stage.

After experimentation with music signals, we have established a default modulation frequency resolution of 0.25 Hz. The modulation frequency resolution affects how many sources are identified in the modulation spectrum, which in turn directly affects the number of resulting separated signals. Also, this resolution may be decreased after each iteration of the framework when using the output, or separated signal, as the input for the next iteration. This could allow for additional degrees of separation at each iteration.

4.2.4.3 Parameter 3: Acoustic Frequency Subband Widths for Reconstruction

In the reconstruction stage (Section 4.2.3), we use the information retrieved about sources in the identification stage to determine which subbands are needed in reconstruction. We

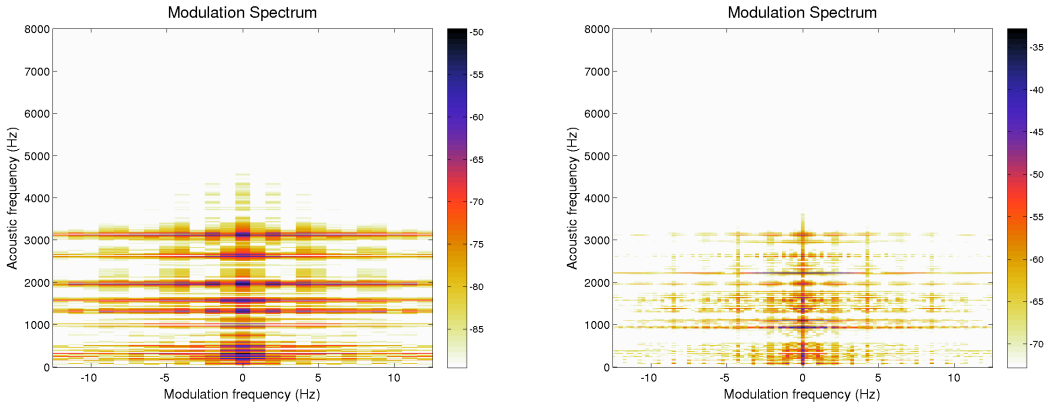


Figure 25: Modulation spectra of the same music signal with modulation-frequency resolution of 1 Hz (left) and 0.25 Hz (right).

may need to revisit the decomposition stage with a different acoustic subband width for demodulation to obtain either narrower or wider modulators to use in reconstruction. Similar to properties of the acoustic subband width parameter for the modulation spectrum (Section 4.2.4.1), the acoustic subband width affects demodulation of the signal in the time domain prior to the reconstruction stage.

Ideally, the acoustic subband width for the reconstruction stage would be the same as whatever width was used for the initial decomposition stage prior to identification. However, we found that by increasing this parameter for the reconstruction stage, we are able to include neighboring frequency subbands for the identified sources, which may increase the quality of the resulting separated signals. Practically, increasing this value may result in the inclusion of previously missing portions of a modulator signal, which may increase sound quality. Sound quality may then be improved since the modulator signal is more in tact. Although, this may be unfavorable if extra, non-target sounds are also included in the larger subband width. As a rule of thumb, the subband width should not be less than it was in the identification stage, because that may eliminate all target sound sources that were originally identified as having the target modulation frequency. Adjusting this parameter value has no affect on the number of resulting separated sources since this information has already been determined in the identification stage.

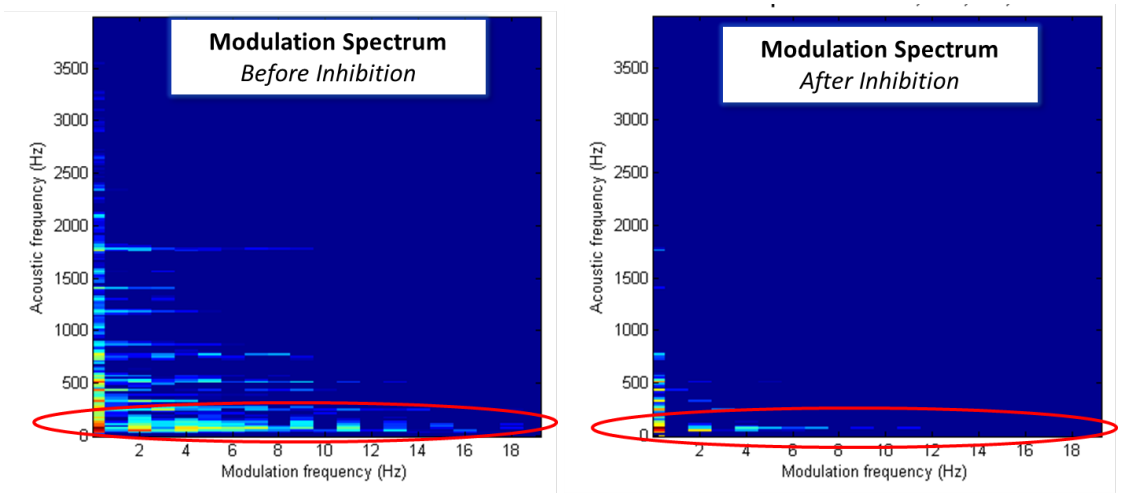


Figure 26: Two modulation spectra of the same signal, one without any thresholding (left) and the other with thresholding (right) to show how structured components may be enhanced and less significant components may be removed prior to undergoing source identification.

4.2.4.4 Parameter 4: Threshold for Modulation Spectral Amplitude

As an inhibition process, a threshold may be set to limit which components of the modulation spectrum are used in the unsupervised source identification algorithm. Since components with the least amplitudes are usually not the focus of the separation analysis, especially since these components may contain non-target sounds or noise, we ignore these components so they will not interfere with analysis. An example is shown in Fig. 26. Our default threshold is set to ignore all components with modulation spectral amplitudes less than 10 percent of the maximum amplitude in the modulation spectrum of the signal. The results of this threshold affect the number of sources that are identified, which in turn directly affect the number of resulting separated signals. If the threshold value is increased, less sources may be identified, which may be helpful for focusing on those components with the highest modulation spectral amplitudes. In contrast, decreasing the threshold may help avoid limiting the weaker components, especially if signals contain only a few sound sources.

Table 3: Calculation of Correlation Coefficients between Modulators with Minimum Threshold (r) for k Subbands

	$m_1(n)$	$m_2(n)$	...	$m_k(n)$
$m_1(n)$	1	$\rho_{1,2} > r?$	$\rho_{1,...} > r?$	$\rho_{1,k} > r?$
$m_2(n)$	$\rho_{2,1} > r?$	1	$\rho_{2,...} > r?$	$\rho_{2,k} > r?$
...	$\rho_{...,1} > r?$	$\rho_{...,2} > r?$	1	$\rho_{...,k} > r?$
$m_k(n)$	$\rho_{k,1} > r?$	$\rho_{k,2} > r?$	$\rho_{k,...} > r?$	1

4.2.4.5 Parameter 5: Threshold for Grouping Time-Domain, Modulator Signals for Reconstruction

Due to the threshold for modulation spectral amplitude (Section 4.2.4.4) in the identification stage, some components associated with sources and their acoustic frequency subbands may be missed. To compensate for this, we group modulators that are the most correlated with each of the identified source's subband modulators. We accomplish this by setting a threshold for correlation coefficients, which is calculated according to Pearson's correlation coefficient, in Eq. 5

$$\rho_{target,j} = \frac{C_{target,j}}{\sqrt{{\sigma_{target}}^2 * {\sigma_j}^2}}, \tag{5}$$

where j is the j-th modulator from $1 \ldots k$ subbands. We calculate the correlation between a target modulator and each of the other modulators along other $k - 1$ subbands in the given signal. A correlation coefficient equal to 0 indicates no correlation while a correlation coefficient equal to 1 is completely correlated, as in the signals are identical. Table 3 demonstrates how modulator signals would be selected for a given target modulator in the leftmost column. Decreasing this value may cause more modulators to be grouped together, possibly creating a "fuller" sound. However, lowering the threshold may also leave room for non-target sounds to enter the reconstruction of the resulting separated signal. Our default threshold for the correlation coefficient for each comparison is 0.7.

4.3 Results and Analysis

We conducted several experiments on data from publicly available datasets including RWC[3], SiSEC[4], TRIOS[5], University of Iowa Music Database[6], and LabRosa[7], which are grouped and discussed in individual case studies along with a listening test. The experimental signals, which were between 5 and 20 seconds long, consisted of 30 custom-made sample signals and clips from over 100 professionally recorded song samples, all sampled at 16 kHz. Example sound files of some of our results may be heard online (see Section 4.4 for link).

4.3.1 Computational Concerns

Prior to discussing the case study results and examples, we briefly interject to discuss runtime of the overall framework. Runtime is dependent on the parameters and, of course, the signal's length and sampling frequency. For our experiments, we used a Linux environment (Ubuntu 12.04) on a desktop computer with an i5 Intel processor and 8 GB of RAM. The framework analyzed and separated a signal in less than a minute for most of our sample signals, regardless of the parameters given. The runtime decreases for one or more of the following parameter options: wider acoustic frequency subbands for modulation spectra used in identification and/or time-domain demodulation used in reconstruction, coarser resolution across the modulation frequency axis in the modulation spectra used in identification, and higher maximum thresholds for the modulation spectral amplitude used in identification and/or the correlation coefficient used in grouping modulators during reconstruction.

[3] https://staff.aist.go.jp/m.goto/RWC-MDB/
[4] http://sisec.wiki.irisa.fr/tiki-index.php?page=Audio+source+separation
[5] http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/27
[6] http://theremin.music.uiowa.edu/MIS.html
[7] http://labrosa.ee.columbia.edu/sounds/

Table 4: Objective Measures for Reconstructed Signals

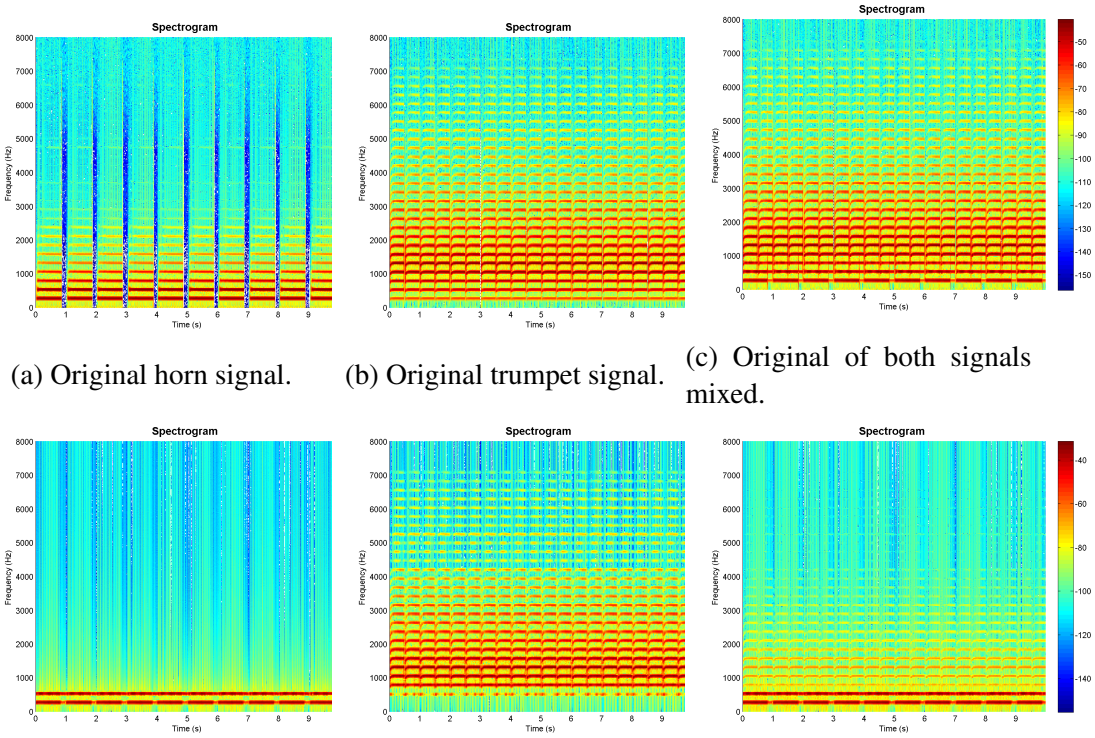
	SDR(dB)	SNR(dB)	WMSD
Source 1 (Horn)	5.80	59.64	4.24
Source 2 (Trumpet)	14.87	68.79	2.47
Sources 1 and 2	17.65	55.98	8.28

4.3.2 Case Studies for Unsupervised Source Separation

In this section, we discuss four case studies and summarize results for each. For our first case study, we performed separation on polyphonic, monaural signals containing harmonic sources without percussive sources. We demonstrate the separation of two harmonic signals that were mixed together. One signal contains a horn note at middle-C repeated once every second, and the other signal is a trumpet note at middle-C repeated every half second. In Figs. 27 and 28, the spectrograms and modulation spectra, respectively, are compared for the original and reconstructed signals. We also calculate the signal-to-distortion-ratio (SDR), signal-to-noise-ratio (SNR), and the weighted modulation spectral dissimilarity (WMSD) measure (Appendix A) in Table 4 using a blind source separation evaluation toolbox for MATLAB.[8] Since our algorithm searches the modulation spectra for harmonics (integer multiples of similar modulation components vertically along a modulation frequency column) and groups similarly shaped modulators, we are able to cluster harmonic sources belonging together even if some components of the sources overlap in pitch (i.e. in the same acoustic frequency subband.) Bass parts, especially in contemporary genres of music, were easiest to identify and separate. We note that the fewer the number of sources in a signal, the easier they were to identify and separate.

Secondly, similar to the example of modulators shown in Fig. 22, we show another example with both percussive and harmonic sources. These usually work well and tend to separate percussive sources from the harmonic melody or accompaniment. We demonstrate this with the "Take 5" trio example from the TRIOS dataset. This 6-second, segmented portion of the song contains drums and piano, which separated nicely, as shown in Figs. 29

[8] http://bass-db.gforge.inria.fr/bss_eval/



(a) Original horn signal. (b) Original trumpet signal. (c) Original of both signals mixed.

(d) Horn signal after separation. (e) Trumpet signal after separation. (f) Both signals modified* and mixed.

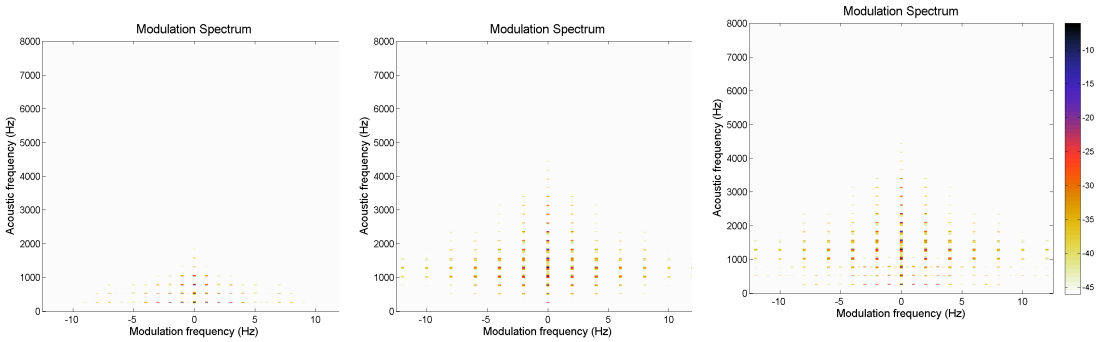
Figure 27: Spectrogram comparisons of original and reconstructed signals after separation or mixing. *In particular, (f) is a spectrogram of a mixed signal containing the two separated signals, but with the trumpet attenuated and the horn amplified.

and 30. Although the SDR values are negative in Table 5, our listening study (Section 4.3.3) resulted in this signal being one of the highest ranked in both separation and quality. This observation suggests that the WMSD measure may be more accurate in measuring perceptual quality.

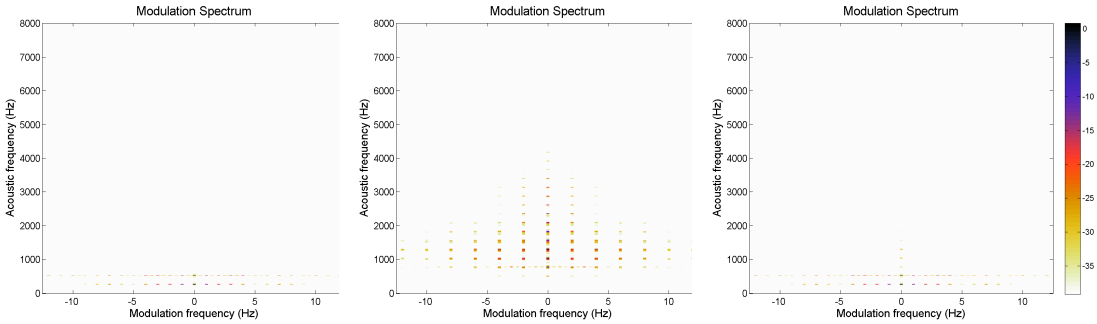
In our third case study, experiments were conducted on signals that contained percussive sources only. Results varied for these types of signals. Some signals were able to be separated while others, particularly the percussive sources that both appear and sound

Table 5: Objective Evaluation for Reconstructed Signals

	SDR(dB)	SNR(dB)	WMSD
Source 1 (Piano)	-5.29	81.85	3.46
Source 2 (Drums)	-13.17	88.28	21.69
Sources 1 and 2	-22.24	81.33	2.79



(a) Original horn signal. (b) Original trumpet signal. (c) Original horn and trumpet signals.

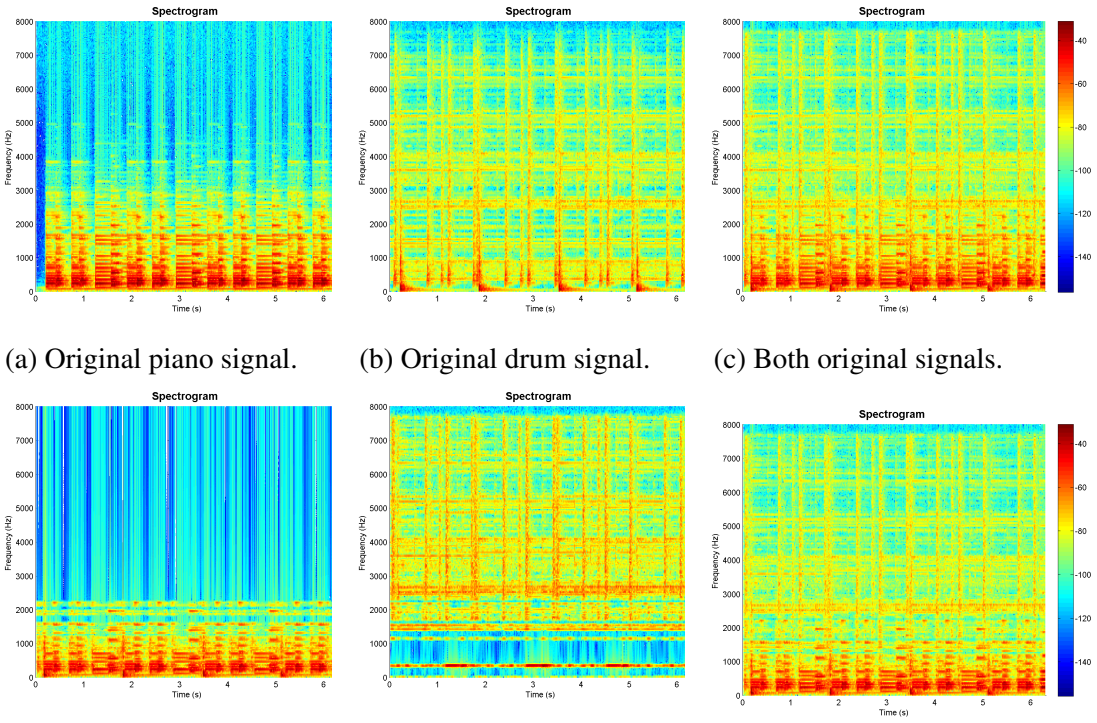


(d) Horn signal after separation. (e) Trumpet signal after separation. (f) Both signals modified* and mixed.

Figure 28: Modulation spectra comparisons of original and reconstructed signals after separation or mixing. *In particular, (f) is a spectrogram of a mixed signal containing the two separated signals, but with the trumpet attenuated and the horn amplified.

"noisy" in the modulation spectra, were more difficult to identify and isolate. These noisy sources include cymbals and other sources that do not necessarily form evenly-spaced harmonics in acoustic frequencies. Some percussive instruments are more easily distinguished than in others (see Fig. 31). However, we observed that the grouping of similar-shaped modulator windows may be able to compensate for some of this lack of distinction. For example, the bass drum and the snare drum have differently shaped modulators, and these modulators may more clearly demonstrate differences in temporal patterns.

Our final case study consisted of experiments with mixed sources that also contained vocals. We note that if a source possesses an unclear temporal pattern, such as with vocals in a song, these sources will not be easily identified. Therefore, vocals, along with other nonstationary sources, are not separated easily using our framework, which searches for



(a) Original piano signal. (b) Original drum signal. (c) Both original signals.

(d) Piano signal after separation. (e) Drum signal after separation. (f) Both separated signals.

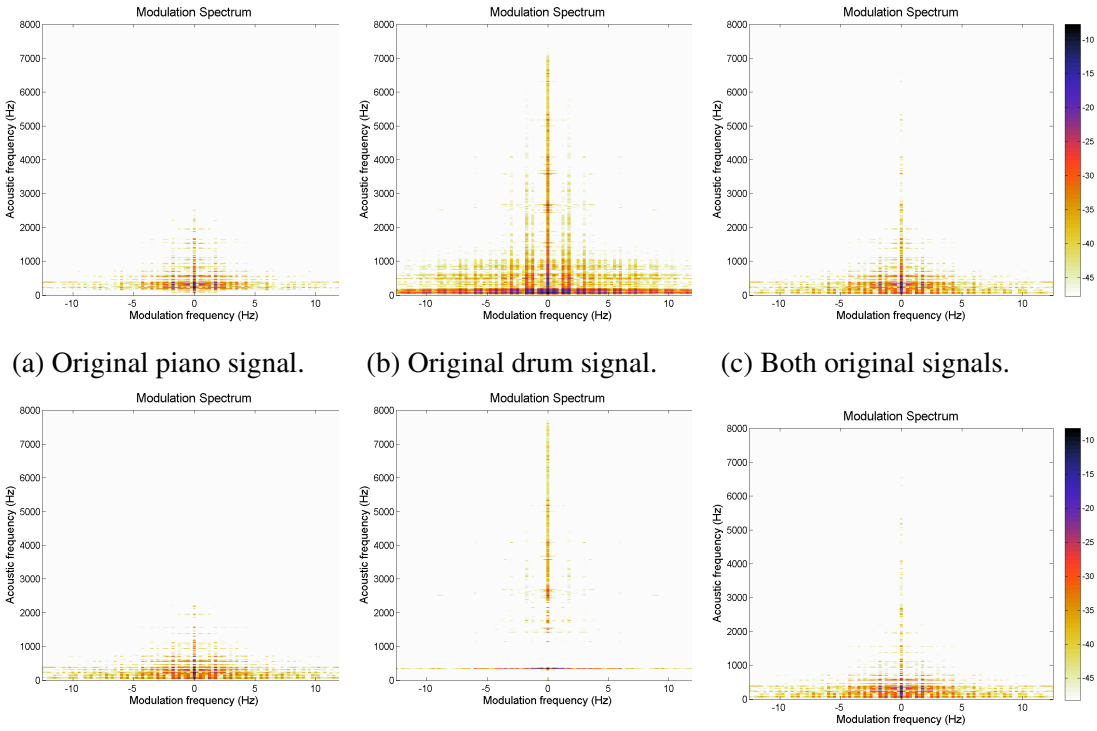
Figure 29: Spectrogram comparisons of original and reconstructed signals after separation or mixing.

sources with temporal patterns. An example is shown in Fig. 32, taken from the LabROSA samples ("Around the World" by ATC). The observation of vocals being difficult to isolate was also reflected in our listening tests (Section 4.3.3).

4.3.3 Listening-Test Study

A listening test was performed with 10 sets of listening samples to compare and rate. This test was estimated to take approximately 10-15 minutes.[9] This study was patterned after the PEASS Listening Test GUI for MATLAB.[10] For each of 10 test sets, the listener was given a specified target sound to listen for, a mixed track sample that contained the specified target sound along with other sounds, and a pair of resulting samples (Sample A and Sample B). These pairs included separation results from different algorithms, default versus adjusted

[9]This study was approved by the Georgia Tech Institutional Review Board.
[10]http://bass-db.gforge.inria.fr/peass/PEASS-ListeningTestGUI.html



(a) Original piano signal.

(b) Original drum signal.

(c) Both original signals.

(d) Piano signal after separation.

(e) Drum signal after separation.

(f) Both separated signals.

Figure 30: Modulation spectra comparisons of original and reconstructed signals after separation or mixing.

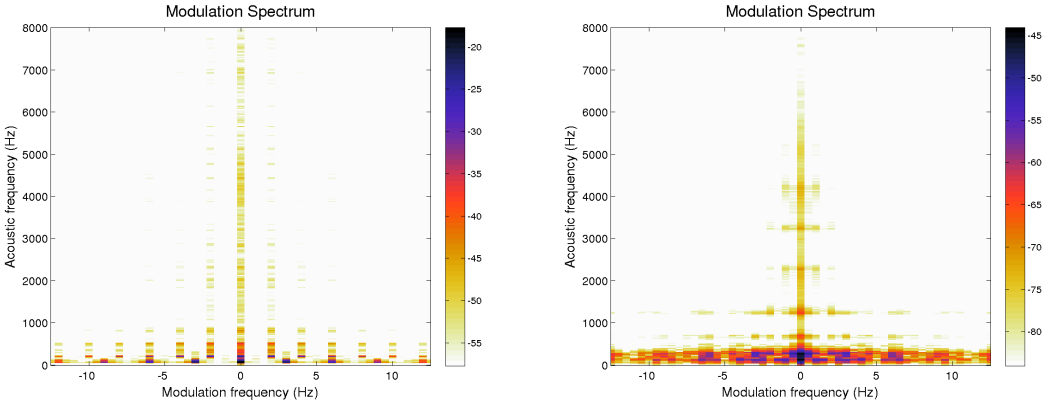


Figure 31: Modulation spectra demonstrating how percussive signals may vary in isolation difficulty during the ID stage of the framework. The left signal contains a bass (2 Hz) and snare (3 Hz) and shows distinguishing source components. The right signal contains several types of percussive sounds and shows less distinguishing source components. Note that "harmonics" of most of the percussive sounds are not evenly spaced in acoustic frequency (unlike harmonic sources).

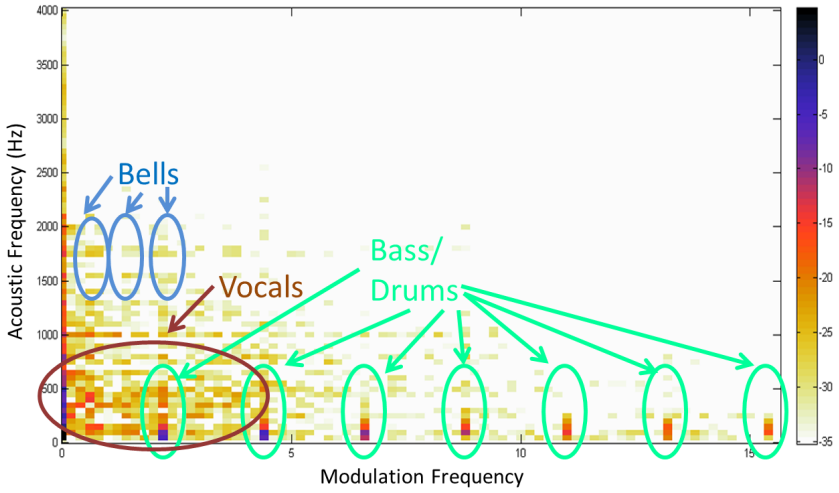


Figure 32: Modulation spectrum demonstrating how difficult vocals are to isolate and recognize patterns in during the ID stage of the framework.

parameters, and *pass* versus *stop* signals. The listeners were then asked to perform two rating tasks for each of the 10 test sets:

1. Rate how well Sample A and Sample B eliminate interfering sounds and extraneous, artificial sounds (i.e. in an attempt to isolate the specified target sound from the mixed track sample).

2. Rate Sample A and Sample B strictly according to sound quality and least amount of distortion (i.e. regardless of the amount of target sounds present).

The first rating task is a subjective measure of our separation results in terms of the presence of extraneous artifacts and interfering sounds. The second rating task is a subjective measure of distortion within our separation results. Rating was done on a scale from 0-100, where 0 is worst and 100 is best, and we later map our analysis to rating labels according to that in Table 6. The listener was asked to make sure that the ratings between pairs of samples were consistent, i.e. if one sample has better quality than another, it should be rated better. They were also allowed to play the samples as many times as they liked.

The samples, shown in Table 7, were chosen to test percussion, vocals, temporal versus non-temporal patterns, harmonics, particular instruments, groups of instruments, etc. Out

Table 6: Rating-Labels and Percentages

Percentage	0-20 %	21-40 %	41-60 %	61-80 %	81-100 %
Label	Low-to-none	Low Moderate	Moderate	High Moderate	High

Table 7: This table describes the data samples used in the listening test. Each sample, about 10-20 seconds long, contained a somewhat stationary segment of the song listed.

Testsets	Song	Origin
1, 5, and 9	Take Five	TRIOS Dataset
2	Jingle Bells	RWC Database
3	Mixed (test_nodrums_liverec_250ms_5cm_mix.wav)	SiSEC Test Dataset
4	Yankee Doodle	RWC Database
6	Eien	RWC Database
7	Mixed (test2_wdrums_inst_mix.wav)	SiSEC Test 2 Dataset
8	Koi	RWC Database
10	Dance of the Sugar Plum Fairies	Tchaikovsky's Nutcracker Suite

of the ten test sets given, five of the test sets compared a separated target track with remaining tracks. Another test set compared modified parameters for a separated target with the default parameters. Two other test sets compared our framework with iCARMs results (including one mixed track that included vocals).[11] The last two test sets compared our framework with SiSEC results (including one track with vocals).[12] [13]

We solicited participation over the course of three weeks from the AUDITORY mailing list[14], the MIR mailing list[15], Georgia Tech's Digital Signal Processing Students mailing list, and other colleagues to gain a diverse pool of participants. 126 people from many different countries in North America, Europe, and Asia participated, as shown in Fig. 33. We used data from the 63 participants who fully completed the listening test and provide demographic information in Figs. 34-36. 12.7% were 18-24 years old, 52.4% were 25-34 years old, 22.2% were 35-50 years old, and 12.7% were age 51 or older. 63.5% were male and 36.5% were female. None indicated any known hearing issues, and the majority

[11] iCARMs results for RWC-MDB-P-2001 No.1 and RWC-MDB-P-2001 No.5, both from the RWC music database were taken from `https://staff.aist.go.jp/k.yoshii/icarm/`

[12] One signal was the SiSEC result from B. Makkiabadi's algorithm: `http://www.onn.nii.ac.jp/sisec13/evaluation_result/UND/submission/bm/Algorithm.txt`

[13] Another signal was the SiSEC result from Adiloglu, Kayser, and Wang's algorithm: `http://www.onn.nii.ac.jp/sisec13/evaluation_result/UND/submission/ob/Algorithm.pdf`

[14] http://www.auditory.org/

[15] http://listes.ircam.fr/wws/info/music-ir

Figure 33: Location representation of listening-test participants.

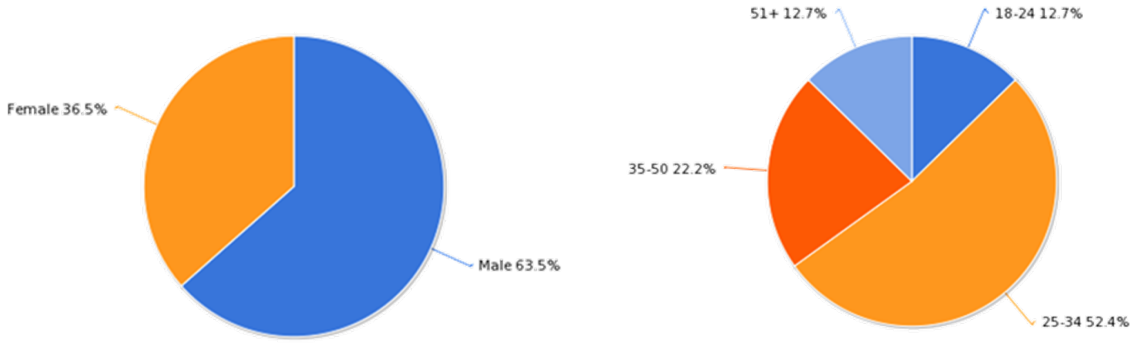


Figure 34: Gender representation (left) and age groups (right) of listening-test participants.

reported they played a musical instrument and/or claimed to have a moderate-to-expert musical background. The majority used a desktop or a laptop computer, with over 80% of them using headphones.

Our listening-test results are presented in Table 8. In all cases where our framework was tested to determine whether or not the target sound was separated from the other sounds, our participants confirmed success, reporting quality level in the moderate-to-high range.

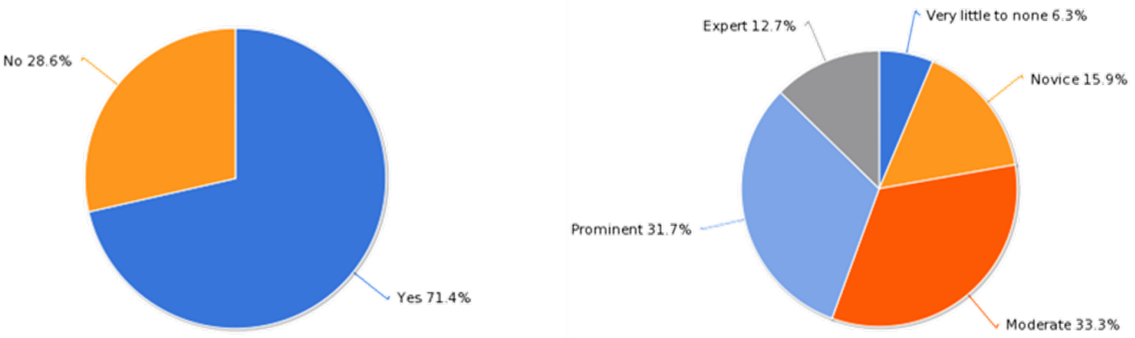


Figure 35: Amount of listening-test participants who play/have played a musical instrument (left) and their self-proclaimed levels of being music enthusiasts (right).

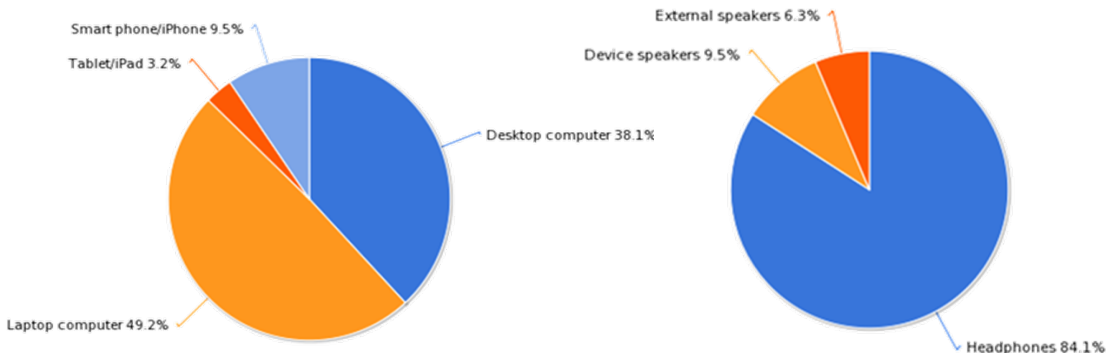


Figure 36: Number of types of devices used by listening-test participants to access the survey (left) and number of audio output devices used by listening-test participants (right).

According to our participants, some of the SiSEC algorithms yielded better quality. This may be expected, given that the SiSEC algorithms required, and were provided, prior information that was not provided to our framework. Nonetheless, our separation performance and sonic quality were rated only slightly worse. In comparisons of our framework with other algorithms involving vocal samples, our framework was considered slightly worse in separation and quality. Overall, our framework's separation of percussive sounds reported separation and quality to be in the moderate-to-high-moderate range. Our modified parameters, consisting adjustments made to default parameters after listening to original results, ranked similar to the default performance in terms of separation performance but demonstrated the ability to yield better sonic quality.

We compared results from iCARMs and our framework, since both methods are unsupervised. For the bass guitar, our framework rated better than iCARMs results in sonic quality and rated much better in terms of separating and eliminating extraneous sounds. For vocal samples, iCARMs rated slightly better than our framework, in terms of both separation ability and sonic quality. Both iCARMs and our framework were rated in the low range in terms of sonic quality for separated vocals. We conjecture that iCARMs may be slightly better for separating some instruments while our framework may be slightly better for others.

Table 8: This table displays results of the listening test comparing pairs of samples in each testset in terms of two rating types: how well the signal separated or contained the target sound(s) and how great the quality of sound (regardless of separation quality). Note: All signals are a result of our framework unless otherwise noted in parentheses.

			Separation Quality		Sound Quality	
Testset	Sample	Signal Details	Avg. %	Std. Dev.	Avg. %	Std. Dev.
1	A	Target:Piano	72.87	19.49	65.42	24.80
1	B	Non-Target	29.92	31.58	37.89	27.96
2	A	Target:Bells	65.00	24.23	63.44	21.13
2	B	Non-Target	48.75	30.82	63.50	23.15
3	A	Target:Male Vocals	47.54	22.48	42.83	23.19
3	B	Target:Male Vocals (SiSEC)	51.58	24.72	66.72	19.89
4	A	Non-Target	34.92	26.92	54.60	22.79
4	B	Target:Drums and Bass	67.34	21.36	55.82	24.70
5	A	Target:Saxophone (Our Optimal Params.)	53.41	21.78	50.82	22.94
5	B	Target:Saxophone (Our Default Params.)	52.42	23.15	39.34	23.11
6	A	Target:Bass (iRCAMs)	39.07	22.84	29.43	20.02
6	B	Target:Bass	59.03	30.77	46.59	26.65
7	A	Target:Crank Sound (SiSEC)	53.79	26.39	45.00	25.77
7	B	Target:Crank Sound	50.95	24.83	48.95	26.73
8	A	Target:Female Vocals (iRCAMs)	41.11	22.58	33.49	21.83
8	B	Target:Female Vocals	36.59	23.06	29.84	21.29
9	A	Target:Snare and Cymbals (Our Optimal Params.)	57.66	24.70	52.82	26.21
9	B	Target:Snare and Cymbals (Our Default Params.)	35.41	34.53	64.43	24.71
10	A	Non-Target	40.16	37.65	73.81	23.80
10	B	Target:Pizzicato Strings	69.27	30.17	63.33	30.46

4.4 Conclusion and Future Work

We have developed a framework for exploiting modulation spectral features for music information retrieval tasks such as unsupervised source separation. Our three-stage framework may be employed in various ways to achieve various tasks and may be used iteratively. For example, a signal that is chosen from the output of the source separation stage may be further treated as a new input signal of the framework. This chapter focused on unsupervised source separation and showed several case studies for various types of sources. A listening study supported the subjective quality of our results in several types of unsupervised source separation. We demonstrate how a set of default parameters may be adjusted, depending on the signal, to achieve desired results. Example sound clips[16] of our results and our listening test[17] are temporarily available online. One can find results of the well-suited signals, such as those that have strong temporal patterns with no significant overlap in both acoustic and modulation frequencies. Also, one can find results of those signals that

[16]http://csip.ece.gatech.edu/?q=student/nashlie-h-sephus

[17]http://edu.surveygizmo.com/s3/1569000/Subjective-Evaluation-for-Sound-Separation-Synthesis-in-Music

are not well-suited for our framework, i.e., those containing non-stationary patterns, vocals, and patterns of different sources that may overlap in both frequency types, such as the bass drum. Since the framework parameters depend on the signal itself, the limitations in results may be attributed to non-optimal parameter settings as well, as discussed in Section 4.2.4.

Future work might integrate other existing methods for recognizing features of vocals, since our framework is not generally geared towards speech or vocals in general. A gradient descent method, or some other optimization technique, for automatically finding an improved set of parameters given initial estimates or defaults would be of interest. Another option would be to use non-uniform subband widths and automatically determine these widths via pre-clustering along acoustic frequency subbands. Other ways might be incorporated to better separate sound sources that overlap in both acoustic and modulation frequencies. Lastly, in addition to our method for grouping correlated modulators, more robust ways of detecting envelopes and smoothing of envelope signals could be developed to improve the framework's ability to distinguish sources.

CHAPTER 5

ADDITIONAL APPLICATION: MODULATION ANALYSIS IN EEG SEIZURE SIGNALS

[1]Epilepsy is a neurological disorder that affects millions worldwide with recurrent seizures that impair quality of life as well as brain function [87]. Many epilepsy patients undergo invasive electrophysiology, or intracranial EEG (iEEG) to pinpoint critical brain area(s) responsible for their seizures, or seizure onset zone (SOZ), by neurologists and epileptologists and likely to result in seizure freedom if surgical removed by a neurosurgeon in a procedure called presurgical evaluation [88]. Standard presurgical evaluation involves board-certified epileptologists visually manually reviewing video and electrophysiology for seizuresif any seizures even occurduring patient hospitalization and monitoring [88]. Consequently, presurgical evaluation is a somewhat subjective process that can have inter-rater variability, although done by trained physicians, and a tedious often challenging process for physicians to manually perform. Thus, the need to objectively pinpoint the SOZ with reliable accuracy, efficiency, and low procedural burden presents an opportunity for novel signal processing techniques to perform the same process as a human EEG reviewer. The modulation spectrum [8] is perhaps a technique that may effectively aid seizure diagnosis.

Recently, a neuroimaging study of temporal lobe epilepsy reported diminished default mode network organization in the same hemisphere as the seizure onset zone [89]. But presently it is unclear how such brain network activity changes at the iEEG level or in neocortical epilepsy. Previous researchers have believed that low frequency rhythms, such as δ or θ oscillations, may couple with high-frequency rhythms (e.g., γ oscillations) for normal or pathological brain function [90–92]. Because the modulation spectrum technique is a straightforward sound approach to quantify network communication via time series

[1]This chapter is modified from *Smart, O. L., Sephus, N. H., & Gross, R. E. (2014). Application of Modulation Spectrum for iEEG Seizure Analysis. In 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (To be published in May 2014 proceedings)* [86].

data, we applied it to a few human epileptic seizures.

Previous research has exploited the amplitude modulation (AM) and frequency modulation (FM) characteristics of the modulation spectrum for primarily audio signal processing (e.g., recognition, filtering) [81, 93]. But some related work has been done in biological or clinical applications [38, 47, 94]. One paper initially sparked interests in bio-inspired modulation analysis for biological applications, particularly for detecting dysphonia in a human speech [38]. Another paper applied an amplitude modulation analysis method on EEG signals for diagnosing Alzheimer's disease [47]. More biomedical research has employed it as a preprocessing step for filtering heart and lung sounds and separating the two for further individual analysis [95]. Thus, application of the modulation spectrum for biomedical time-series analysis has gained growing interest.

Surprisingly, the modulation spectrum algorithm has not yet been applied to epilepsy time-series signals. In fact, in the field of signal processing techniques for epilepsy iEEG or EEG, cross-frequency coupling and modulation index methods are just starting to grow in application [94, 96–98] with most approaches focused on one of two well-known joint-frequency modulatory signal representations [99, 100] that though potent in use have some drawbacks [101] due to their stochastic nature. In this *in silico* study, we investigated whether modulation spectrum measures varied across electrode location (i.e., SOZ, non-SOZ) or seizure state changes (i.e., preictal, ictal, postictal). We found that the modulation spectrum could usefully quantify changes in seizure state and location but that these results may be patient-specific despite a general trend in the measurements.

5.1 Methods

5.1.1 Epilepsy Patients and Invasive Brain Signal Data

Brain signals were obtained from four medically refractory epilepsy patients who underwent long-term simultaneous time-locked video and IEEG recordings for presurgical evaluation, which entailed the surgical implantation of platinum-iridium brain electrodes (Ad-Tech Medical Instrument Corporation, Racine, WI, USA) that were later extra-operatively

connected to a 128-channel data-acquisition system (Xltek, Oakville, Ontario, CA). The acquisition system stored signal data at 12 bits per sample and at 500 samples per second (Patients B-D) or 1000 samples per second (Patient A) before downsampling to 500 samples per second in a referential montage. Each patient provided permission for the access and analysis of their data via an informed consent process (Emory University, IRB00027074).

During their hospitalization, each patient experienced at least one behavioral seizure with corresponding electrographic discharge (i.e., iEEG seizure). An epileptologist annotated SOZ electrodes and seizure onset and offset times per seizure. For our analyses (Sections B-D), we clipped iEEG signals for a single randomly selected seizure with clear clinical annotations per patient. Each iEEG seizure (i.e., ictal) clip ranged from approximately 5 minutes before the seizure onset timestamp (i.e., preictal state) and 5 minutes after the seizure offset timestamp (i.e., postictal state). The seizure (i.e., ictal state) duration varied (27-105 seconds) across the four patients.

5.1.2 Modulation Spectrum Theory

The modulation spectrum was developed as an analysis tool to demonstrate the presence of modulation in a signal. In a canonical sense, modulation refers to a slow-varying envelope (i.e., a modulator) that modifies a fast-varying higher-frequency wave (i.e., a carrier); meanwhile the modulation spectrum represents the jointly coupled modulator-carrier relationship of a signal as a quantifiable measure that varies across carrier frequencies versus modulation frequencies. Mathematically, the modulation spectrum may be described as a discrete short-time modulation transform (DSTMT) of a signal [1], as reflected in Eqs. 1-3

We used the modulation spectrum, in particular this specific implementation [8], as a cross-frequency coupling technique rather than other known techniques [99–101] given its very computationally efficient, mathematically sound, statistically deterministic, and easy-to-understand approach. More details on the approach may be found in [1, 7, 81].

5.1.3 Signal Processing with Modulation Spectrum

Using our custom and the open-source MATLAB algorithms, we performed modulation spectrum analysis for each of the four patients' seizure clips in the following manner: (a) we arbitrarily selected a *w*-second window of multi-electrode data for each seizure state (i.e., preictal, ictal, postictal), where *w* equaled the seizure duration and the windows did not overlap in time; (b) we applied de-trending to remove any baseline drift and filtering to remove any line noise (i.e., 60 Hz and harmonics) for each referential iEEG signal; (c) we computed the modulation spectrum for each *w*-second epoch per signal; (d) we considered seven canonical clinically relevant iEEG spectral bands of interest (i.e., δ (1-4 Hz), θ (4-8 Hz), α (8-12 Hz), β (12-30 Hz), γ_1 (30-50 Hz), γ_2 (50-80 Hz), and γ_3 (80-249 Hz)); and (e) we discretized the entire modulation spectrum into 49 major cross-frequency bins for all combinations of the seven spectral bands (Fig. 37), where each cross-frequency bin corresponded to the maximum modulation spectrum value over all frequencies spanning the bin and $bin_{i,j}$ quantified the i-th carrier frequency band and j-th modulation frequency band. We used a $bandwidth_j$:$bandwidth_i$ convention to describe the corresponding physiological iEEG spectral bands as a *modulator*:*carrier* coupling pair.

We used the same parameters for each modulation spectrum computation: the Hilbert Transform for the envelope detection stage (i.e., Hilbert demodulation method [1] discussed in Section 3.2.1), a rectangular window for a Fast Fourier Transform spectral analysis; non-uniform pass-subband widths in the carrier frequency filter-bank (y-axis of the modulation spectrum) so that each passband coincided with the seven iEEG spectral bands; a uniform 1 Hz passband width for the modulator frequency filter-bank (x-axis of the modulation spectrum) from 1-249 Hz; a cubic spline interpolation so that the full bandwidth range of the modulator frequency (1-249 Hz) possessed actually 210 total frequency points (30 points per iEEG spectral band) so that each modulator bandwidth had equitable spectral representation before reduction to a single data point via calculating the maximum modulation spectrum value within that band.

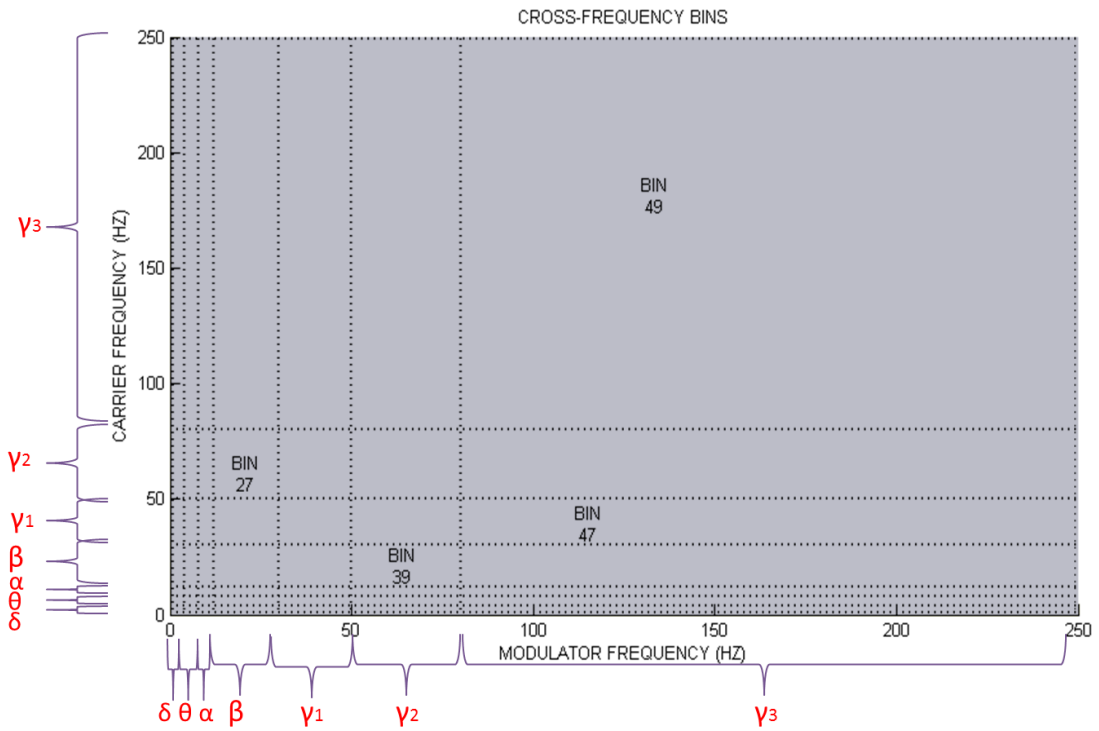


Figure 37: We converted each full-bandwith modulation spectrum result into 49 discrete cross-frequency bins, each bin with an associated index and nomenclature for statistical analyses. For instance, "bin 27" and "bin 49" respectively represented the maximum β:γ_2 and maximum γ_3:γ_3 modulation values as *modulator:carrier* (x-axis:y-axis) cross-frequency signal relationships. A MID plot may be computed using two binned modulation spectrum, each one representing a statistical group for comparison (see Section 5.1.4).

5.1.4 Statistics

We posed two main statistical hypothesis tests for each of the computed 49 cross-frequency measures per patient: (1) equal modulation values in the SOZ versus the non-SOZ (NSOZ) signals for each the preictal, ictal, and postictal epochs; and (2) equal modulation values across preictal, ictal, and postictal epochs in each the SOZ and NSOZ electrodes. We used each of the 49 joint-frequency measures of interest as a dependent variable (e.g., θ:γ_1, δ:α), electrode location (i.e., SOZ or NSOZ) based on clinical annotations as one independent variable, and seizure state as the other independent variable. Each analyzed electrode, therefore, contributed to a statistical sample of modulation measurements per cross-frequency bin.

We used the following non-parametric statistics to test each primary hypothesis: Kruskal-Wallis (KW) one-factor analysis of variance (ANOVA) for an omnibus test and Wilcoxon-Mann-Whitney (WMW) two-sample test for multiple comparisons with the same Bonferroni correction of the standard significance level ($p < 0.05$) for both hypotheses ($p_{KW} < 0.05/49$ and $p_{WMW} < p_{KW}/6 = 0.00017$). In the case of main hypothesis #1, for which the independent variable only had two levels, the WMW test equaled the KW test. For each MWM test, we computed Cohen's d and Hedges's g *effect size* statistics but only report g here.

To compactly illustrate the results of the hypothesis testing for all 49 combinations of cross-frequency bins, we developed a "modulation index difference" (MID) plot. For the MID plot, the x-axis and y-axis represented the seven modulator and seven carrier bands respectively and the z-axis (i.e., a three-level color code) represented the difference in median modulation index across electrodes between two selected levels of an independent variable (e.g., SOZ vs. NSOZ, ictal vs. postictal) weighted by *effect size* (i.e., Hedges's g value) and p-value according to Eq. 6

$$sgn((m_{level_1} - m_{level_2}) * p_{flag} * g_{flag}) \tag{6}$$

where $p_{flag} = 0$ if $p > 0.00017$ (i.e., not a statistically significant result) and $p_{flag} = 1$ otherwise; g is the Hedges's g value with $g_{flag} = 0$ if $g < \tau$ and $g_{flag} = 1$ otherwise; τ is a predefined threshold (e.g., 0.80) and sgn is the signum function. We calculated MID plots with $\tau = 0.80$ (i.e., high *effect size*) then $\tau = 0.30$ (i.e., low *effect size*) for all patients to briefly examine the effects of adjusting this parameter on findings. We did not aim to optimize τ in this work, but aim to so do in future research.

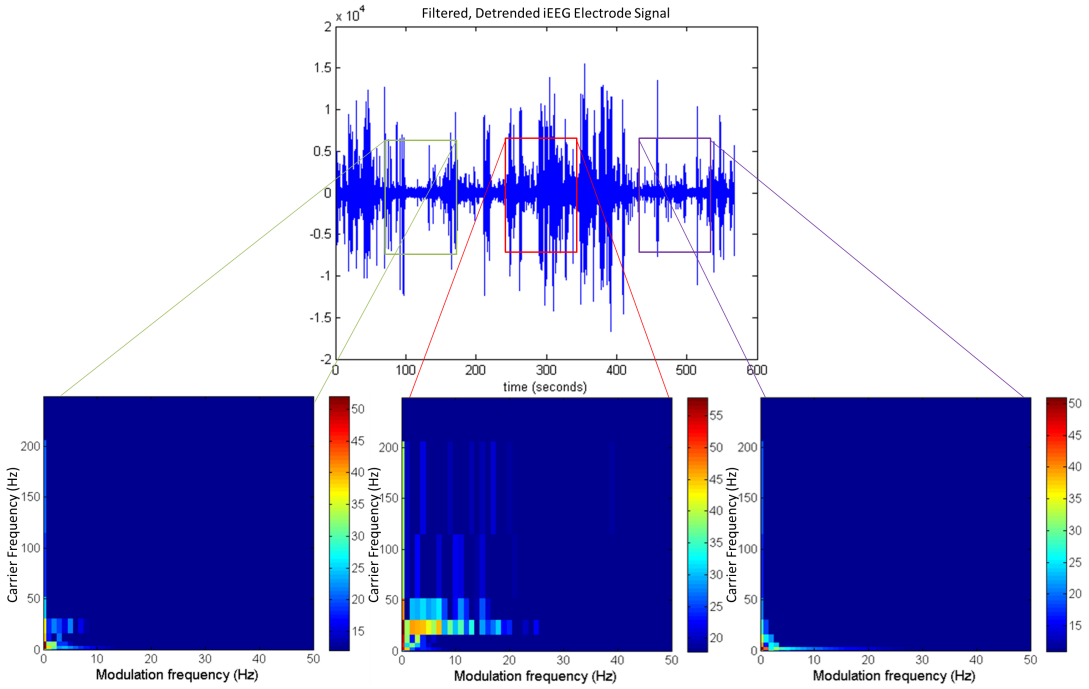


Figure 38: For Patient A (captured here), we analyzed each iEEG electrode by bandpass-filtering (1.0-249.0 Hz) and de-trending its signal (top panel) before computing its modulation spectrum (bottom panels) for preictal (top panel: green box; bottom panel: left plot), ictal (top panel: red box; bottom panel: middle plot), and postictal (top panel: purple box; bottom panel: right plot) epochs. Relatively high modulation was represented by reddish colors while relatively low modulation was represented by bluish colors, where for visualization purposes here we transformed the raw modulation indices to a log-scale. From this type of analysis, we extracted the 49 cross-frequency modulation values per seizure state per electrode for statistical analyses. For this electrode and this patient, we noticed low cross-frequency modulation in preictal and postictal intervals when contrasted with the ictal interval.

5.2 Experimental Results

In preliminary observations, changes were obvious in the modulation spectra across the three seizure states in many electrodes. For instance, we noticed transitions from broad-spectrum coupling in the preictal state to focus-spectrum coupling in the ictal state to even broader-spectrum coupling in the postictal state of an electrode for Patient A (e.g., Fig. 38); we noticed fluctuations in the magnitude and spectral broadness of modulation measures from preictal to ictal and postictal states in Patient C (e.g., Fig. 39). We then succeeded these preliminary observations with two types of experiments, which we discuss in the remainder of this section.

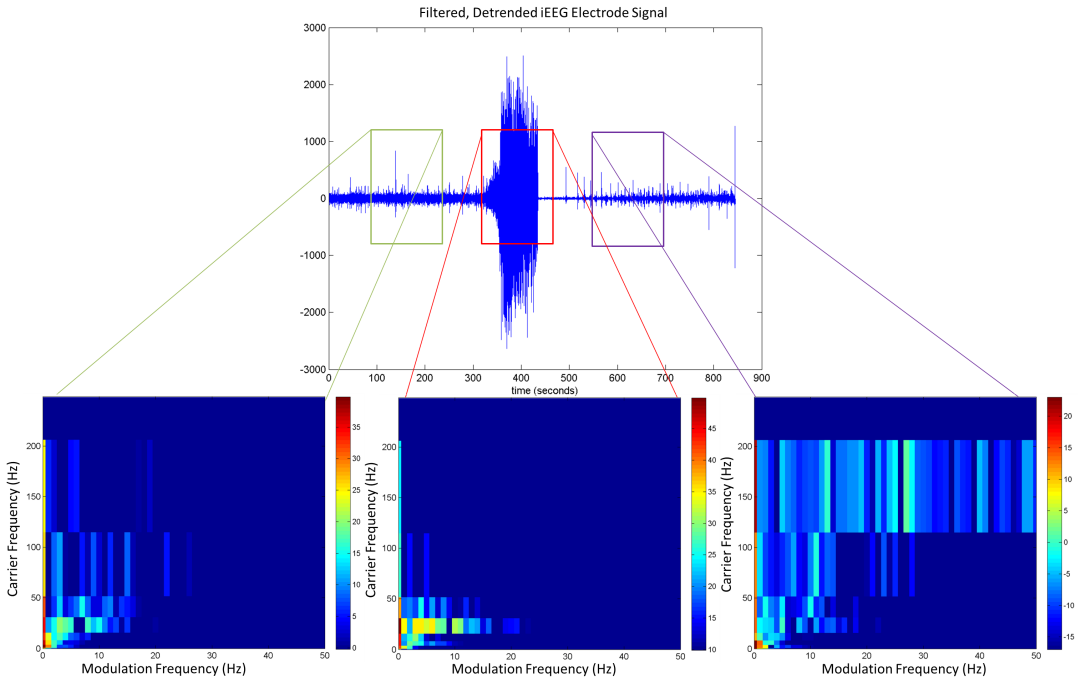


Figure 39: For Patient C (captured here), as with all patients, we performed the same signal processing analysis as with Patient A (Fig. 38). Sometimes we observed different modulation values when juxtaposing preictal, ictal, and postictal epochs across patients and electrodes. For this electrode and this patient, we noticed cross-frequency modulation throughout preictal, ictal, and postictal intervals but the modulation magnitude overall appeared higher and concentrated in less broad spectral ranges in the interval phase versus the other two time intervals.

For our first experiment, we investigated whether each patient displayed differences in cross-frequency coupling for their brain signals outside vs. inside their SOZ. With $g = 0.80$, we found generally no statistically significant differences in NSOZ vs. SOZ coupling across patients and states except for the following 23 of 588 (~4%) total cross-frequency bins (Fig. 40): lower preictal θ:θ coupling in NSOZ than SOZ for Patient D (Fig. 40: D1); higher NSOZ than SOZ coupling of the α:θ, β:θ, γ_1:θ, γ_2:θ, γ_3:θ, θ:β, δ:γ_2, θ:γ_2, α:γ_2, β:γ_2, γ_2:γ_2, and γ_3:γ_2, bands for Patient A (Fig. 40: A2) during seizure; higher NSOZ than SOZ ictal γ_3:γ_3 coupling for Patient C (Fig. 40: C2); lower NSOZ than SOZ coupling of the α:θ, β:θ, γ_1:θ, γ_2:θ, γ_3:θ, δ:β, and α:β bands during seizure for Patient D (Fig. 40: D2); and higher postictal δ:δ coupling and ictal α:θ coupling in the NSOZ than SOZ for Patient A (Fig. 40: A3). However, 20 of the 23 bins (~87%) all occurred for the ictal state,

meaning that Patient B notwithstanding (Fig. 40: B2) patient-specific modulation spectrum measures could distinguish the SOZ and NSOZ electrodes during seizures.

For our second experiment, we investigated whether each patient displayed any differences in cross-frequency coupling brain signal measures when comparing different time intervals relative to seizure onset. With g=0.80, we compared SOZ coupling (Fig. 41) during ictal vs. preictal (Fig. 41: A1-D1), posictal vs. ictal (Fig. 41: A2-D2), and postictal vs. preictal (Fig. 41: A3-D3) iEEG epochs, observing three primary findings: (1) lower modulation during seizure than before seizure for all four patients (Fig. 41: A1-D1); (2) higher modulation after seizure than during seizure for carrier frequencies mainly above β with various modulator frequencies while lower modulation after seizure than during seizure for carrier frequencies mainly below β with various modulator frequencies in Patient A (Fig. 41: A2), higher modulation after seizure than during seizure for only three cross-frequency bins of Patient B (Fig. 41: B2), lower modulation after seizure than during seizure for Patient C (Fig. 41: C2), and higher modulation after seizure than during seizure for various cross-frequency pairings in Patient D (Fig. 41: D2); and (3) lower modulation after seizure than before seizure for various *modulator*:*carrier* pairs in Patients A, C, and D while only two for Patient B (Fig. 41: A3-D3).

We repeated each experiment with a lower *effect size* threshold ($g = 0.30$). We noted that a lower threshold made the results more sensitive to statistically significant differences, revealing some differences in the new analysis (e.g., Fig. 42) that we did not observe in the prior analysis (e.g., Fig. 41). For the first experiment, we noticed more statistically significant differences between NSOZ and SOZ coupling across patients and states: 90/588 (~16%) total cross-frequency bins with 52/90 (~58%) corresponding to the ictal state. For the second experiment, the lower threshold did not considerably alter the observed patterns from the high-threshold analysis although Patient B results varied more (c.f., Fig. 41 and Fig. 42: B) than the other patients.

5.3 Discussion

This study investigated the utility of the modulation spectrum technique, which is a method typically used in speech signal processing although used recently for noninvasive EEG from Alzheimer's patients, for the analysis of invasive EEG from epilepsy patients. To our knowledge, we have contributed the first application of this particular cross-frequency coupling signal processing approach to epilepsy iEEG data. Furthermore, we presented a novel approach to quantitatively identify both time intervals of seizure activity and brain regions conceivably involved in seizure-genesis. Overall we observed patient-specific changes in iEEG modulation spectra: (1) either lower (Patients A and C) or higher (Patient D) coupling in SOZ than NSOZ during seizures; and (2) lower SOZ coupling when a seizure occurs that apparently maintains a postictal effects (Fig. 41: A3-D3).

Regarding the limitations of this work, we recognized the need to more fully examine the effect of the threshold τ on results via this method. We hypothesized from these early observations, however, that some patients may display gradual changes and others more drastic chances in ensuing MID plots as τ varies. Since this threshold affects the interpretation of these analyses, future work must examine the parameter using ROC curves to discern a "cutoff point" for reliable vs. "noise" MID plots. Secondly, we recognized that this method should be further studied by applying it to more seizures per patient and more patients as well as less arbitrary sampling of preictal and postictal iEEG epochs. Thirdly, the method can be applied to iEEG sampled at higher bandwidths since higher γ activity ($\geq$ 250 Hz) has been suggested as a biomarker for SOZ identification. Also, we can examine the same methods for scalp EEG or another brain electrophysiology modality from epilepsy patients. Lastly, our findings have potential translation to a pattern classification framework using coupling measures (or transformed measures, e.g., via ICA or PCA) as extracted features for a machine learning algorithm to discern (a) SOZ vs. NSOZ areas or (b) preictal, ictal, vs. postictal states; but we have begun research on this application [102].

5.4 Conclusion

The modulation spectrum technique revealed that various cross-frequency coupling measurements captured the effects of seizures on brain network activity regardless of whether the electrodes were inside or outside the putative clinical seizure onset zone and that for certain cross-frequency bins modulation measures discriminated electrodes in the SOZ from electrodes not in the SOZ during the seizure state but for a few patients and adjustable *effect size* thresholds. Such findings provide promising potential for the modulation spectrum and measures derived from this analysis for the classification of seizure states and possibly the distinction of SOZ electrodes for diagnostic purposes using iEEG. Our findings can be adapted for a pattern classification method.

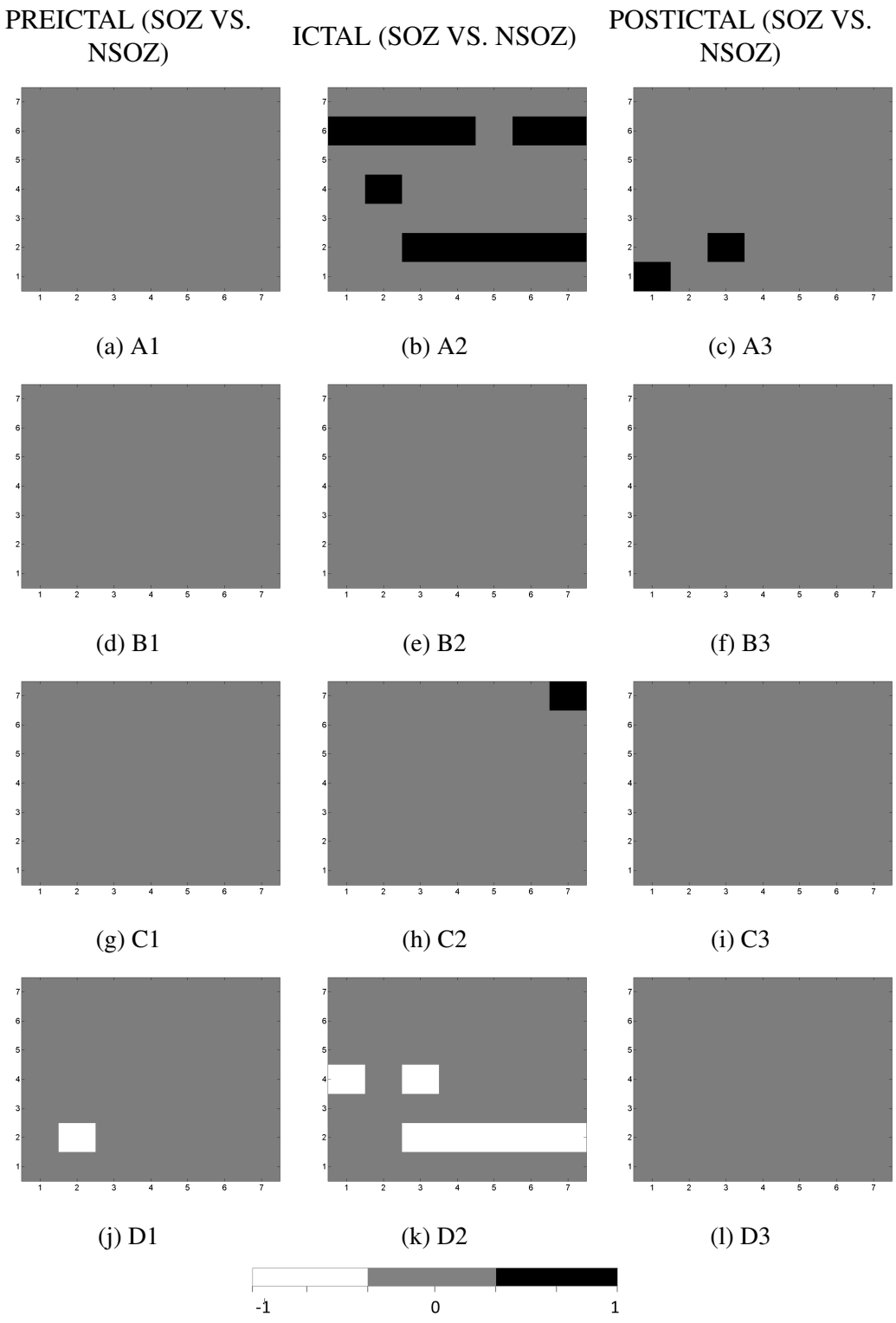


Figure 40: For Patients A-D (rows) we computed MID plots ($g \geq 0.80$) for differences between SOZ and NSOZ modulation spectrum values in each the preictal (first column), ictal (second column), and postictal (last column) time intervals. For all MID plots, the tick marks 1-7 of both the x-axes and y-axes correspond to δ, θ, α, β, γ_1, γ_2, and γ_3 bandwidths respectively; while the pixel colors represent higher coupling in NSOZ than in SOZ (black), no difference in coupling between NSOZ and SOZ coupling (grey), and lower coupling in NSOZ than in SOZ (white).

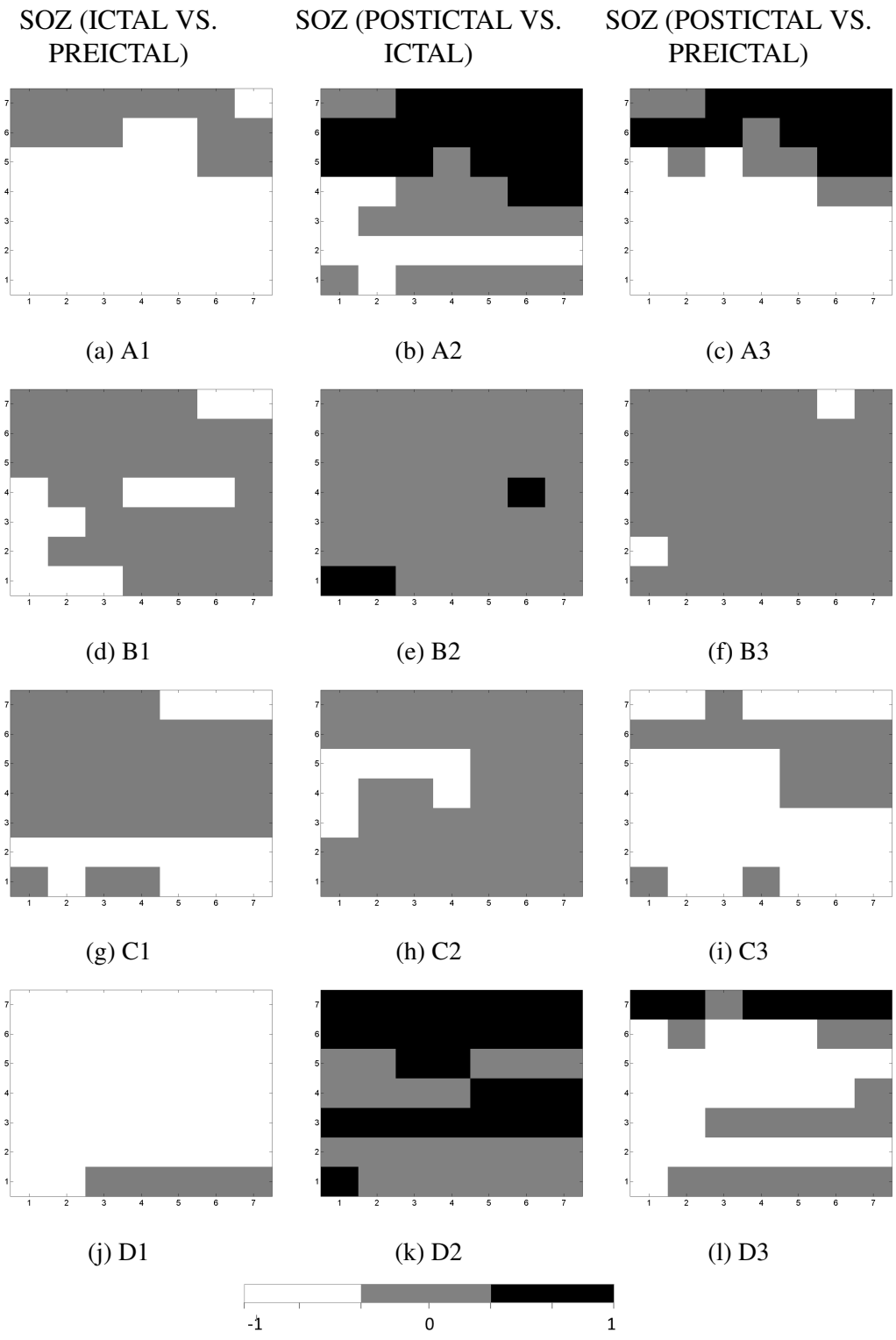


Figure 41: For Patients A-D (rows) we computed MID plots ($g \geq 0.80$) for comparing ictal vs. preictal (first column), postictal vs. ictal (second column), and postictal vs. preictal (third column) coupling in only the SOZ for all plots. For all MID plots, the tick marks 1-7 of both the x-axes and y-axes corresponded to δ, θ, α, β, γ_1, γ_2, and γ_3 bandwidths respectively. The pixel colors per plot in each the ictal vs. preictal, postictal vs. ictal, and postictal vs. ictal comparisons represent higher coupling (black) for a given time interval (e.g., ictal) than its preceding time interval (e.g., preictal), no difference in coupling (grey) between time intervals, and lower coupling (white) for a given time interval than its preceding time interval.

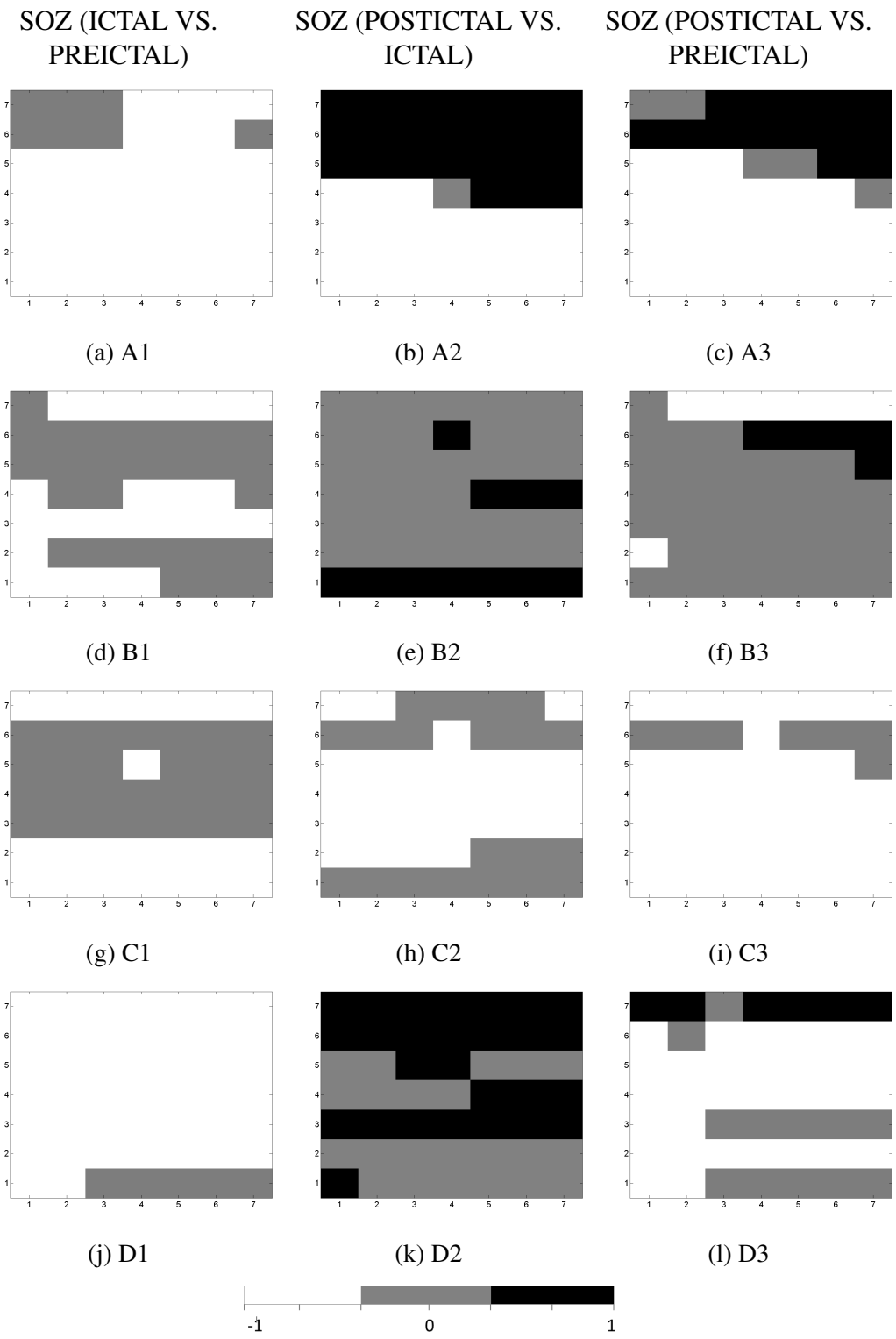


Figure 42: For Patients A-D, we repeated the MID plot analysis in Fig. 41 but with a lower *effect size* threshold ($g \geq 0.30$).

CHAPTER 6

CONCLUSION

In this thesis, we have developed a framework for extracting *modulation spectral features* from music and other signals based on an existing modulation filterbank framework originally employed for speech signals. We have applied this framework to several applications, including unsupervised source identification, unsupervised source separation, and analysis of amplitude modulation (AM) and frequency modulation (FM). The objective of the unsupervised identification method is to automatically identify distinct sources of varying modulation content, a process that is currently manual and requires prior information about sources. The objective of the unsupervised source separation is to blindly separate sources in periodic segments of signals with varying modulation content, or temporal patterns. Other applications presented in this thesis include vibrato analysis and modulation analysis of EEG seizure signals. A summary of our most significant contributions are as follows:

- published extensive literature survey on benefits, challenges, and new directions of exploiting general modulation features in music data mining tasks.
- revisited the modulation transform and modulation spectra theory and drew additional conclusions, theory, and concepts in both short-time and long-term analysis.
  - revealed how FM may be interpreted in the "AM-dominated" modulation spectra.
  - noted how variations in a signal's model may cause modulation frequency to be interpreted differently, especially when using Hilbert demodulation for modulation spectra.
  - emphasized how Hilbert demodulation may be beneficial for music data mining tasks.

- developed a framework for exploiting modulation spectral features in music.
  - performed a parameter study and documented intuitive ways to adjust parameters, which led to an optimal, default set of parameters in modulation spectra for music analysis.
  - developed a new algorithm for unsupervised source identification for segments of music that are somewhat periodic (motivated by HPSS and a modulation filterbank framework intended for speech).
  - developed a new approach for "cleaner" modulation filtering by grouping correlated modulators in the time-domain.
  - developed a new method for unsupervised source separation for signals with strong temporal patterns.
- introduced a WMSD measure to compare signals' modulation spectra directly, i.e. without returning to the traditional frequency domain or time domain.
  - demonstrated feasibility of weights that can be used to skew comparisons in either acoustic frequency or modulation frequency.
- introduced feasibility of a new method for analyzing, comparing, removing, synthesizing, and copying vibrato.
- developed an experimental setup for using non-uniform subband widths along both acoustic and modulation frequency axes in modulation spectra for performing cross-frequency coupling analysis on various states of EEG epileptic seizure signals.

For future work, the MATLAB code used for the framework may be ported to another platform to integrate with open-source music editing/preprocessing software, such as Audacity or Pure Data. Modulation spectra along with the WMSD measure may be useful tools for measuring similarity between modulation features in music. An example application would be providing a quantitative measure for comparing various levels of vibrato

to aid in self-taught or online music lessons [103]. For the framework, one might integrate other existing methods of source identification or source separation to better adapt to a broader range of music signals. This work may incorporate a hybrid coherent and incoherent method for music or an improved envelope detection method dependent on the signal type. As mentioned earlier, a method may be designed to automatically select the best set of framework parameters for a given signal instead of initially using the defaults. Modulation spectral features may be exploited further with psychoacoustic experiments relating modulation analysis to human perception. Referring to FM features and modulation spectra in general, we hope our tutorials, experiments, and results presented in this thesis will provide more intuitive uses and broader applications.

APPENDIX A

WEIGHTED MODULATION SPECTRAL DISSIMILARITY (WMSD) MEASURE FOR MODULATION SPECTRA COMPARISONS

For assessment of tasks performed in the modulation spectral domain, most researchers' focus has been placed on transforming results from this domain back to the time or traditional frequency domains via reconstruction and synthesis. Several metrics are available for quantitative comparison of signals in the time and traditional frequency domains (discussed further in Section A.1). However, we notice a lack in standard assessment methods for comparing signals directly in the modulation spectral domain. For instance, some tasks may not require transformation back to the time or traditional frequency domains. One such task would be automating the evaluation of a music student's performance by comparing modulation spectra of the student's notes with that of a music instructor [103]. For these purposes, a standard dissimilarity measure for modulation spectra, especially of speech and music signals, would be useful.

For comparing signals directly in the modulation spectral domain, we define a *weighted modulation spectral dissimilarity* (WMSD) measure with weighting for both acoustic and modulation frequencies. We calculate the WMSD measure between two modulation spectra, i.e., WMSD(P_c,P_n), in Eq. 7:

$$\left(\frac{10}{M} * \sum_{m=1}^{M} \frac{\sum_{k=1}^{K} W(k,m) \left[\log_{10} \frac{\tilde{P}_c(k,m)^2}{\tilde{P}_n(k,m)^2} \right]^2}{\sum_{k=1}^{K} W(k,m)} \right)^{\frac{1}{2}}, \tag{7}$$

$$\tilde{P}_c(k,m) = \|P_c(k,m)\|, \tag{8}$$

$$\tilde{P}_n(k,m) = \|P_n(k,m)\|, \tag{9}$$

$$W(k,m) = \tilde{P}_c(k,m)^w, w \geq 0, \tag{10}$$

where P_c and P_n are the modulation spectra of the clean, or original, signal and the noisy, or modified, signal, respectively. Both modulation spectra must have the same dimensions.

The integer M is the total number of modulation frequencies (Hz), or column bins, in the modulation spectrum and integer K is the total number of acoustic frequency subbands (Hz), or row bins, in the modulation spectrum. Since these bins may be wide or narrow in resolution, parameters of the modulation spectra may be used to group modulation features in neighboring acoustic and/or modulation frequency bands. Appropriate choices for these parameters enable signals at different time-window alignments, various acoustic pitches, as well as different tempos to appear almost identical in their modulation spectra's representations (see [81]).

Equations 8 and 9 set $\tilde{P}_c(k,m)$ and $\tilde{P}_n(k,m)$ equal to the magnitude of each of the elements of the P_c and $P_n(k,m)$ matrices, respectively, which may be complex values initially. These equations may be modified by relatively scaling the elements from 0 to 1 by dividing each element by the maximum element in each respective matrix. This scaling places less emphasis on the intensity of each signals' modulation components and focuses more on their locations in the modulation spectrum. This could be useful for comparing modulation features of similar signals with different volume levels.

As shown in Eq. 10, $W(k,m)$ are a set of positive weights that may be used to place emphasis on more critical bins along either the modulation frequency axis or the acoustic frequency axis. The weights are the most significant for measuring perceptual quality. The greater the weights, the more emphasis is placed on rhythm when performing long-term analysis. Also, if noise-invariant comparisons of signals are desired (where the noise has little to no modulation characteristics, i.e., random noise), then smaller weights at or near the 0 Hz bin in modulation frequency may enable inhibition of random noise to have less of an effect in the WMSD measure. For our preliminary experiments, we chose $w = 2$, since greater weights generally provide a greater scale of dissimilarity. Psychoacoustic researchers have studied perceptual relationship between the human auditory system and changes modulation [19]. Similar research and experiments with WMSD would lead to greater intuition about systematically choosing weights.

A.1 Background and Related Work

Several metrics and measures of the time and traditional frequency domains have been studied for assessing results of algorithms involving noise removal, signal enhancement, and source separation. Some include the signal-to-noise ratio (SNR) along with its variations, such as segmental SNR and frequency weighted segmental SNR [104]. The segmental SNR is essentially segmented across time frames and averaged together while the frequency-weighted, segmental SNR is weighted across critical frequency bands of the segments and compares short-time energies of signals. The weighted-slope spectral (WSS) measure introduces weighted slopes of the frequency bands in the traditional frequency spectrum. Also, the perceptual evaluation of speech quality (PESQ) is used to gain a more perceptual yet quantitative assessment of quality versus the purely objective SNR. Evaluations of these standard measures have been performed to demonstrate which performed best for various applications [105] [106]. In a study on objective quality measures, the measures with weighted frequencies were found to possess the highest correlation coefficients with subjective quality [104].

For traditional spectral comparisons, features such as MFCCs, spectral flux, spectral centroid, and spectral dissonance are commonly used [9] and have been evaluated as similarity measures in music especially [107]. The most notable measure for comparing modulation spectra would be the modulation spectral contrast (MSC), which is the difference between the modulation spectral peak (MSP) and the modulation spectral valley (MSV) of each logarithmically spaced modulation subband, used for long-term signal analysis in a music genre classification study [14]. Specifically with modulation filtering, assessment techniques have been performed after reconstructing the signal back to the time domain. A study on the analysis of modulation filtering results describes three assessment techniques [80]. The first two techniques calculate distance between the STFT and the modified STFT of a signal via an average squared error and an average least squared error. The third technique was the effective modulation filter response, which is an average ratio of the

original signal's modulation spectrum with the modified signal's modulation spectrum [77]. This technique was used to compare the magnitude responses of modulation frequencies between the two signals.

Another study on separation of heart and lung sounds using modulation filtering evaluates results by comparing power spectral densities (in acoustic frequency) using a log-spectral distance [95]. An iterative algorithm for implementation of improved modulation filtering for speech signals minimizes the squared error of the filtered and the desired modulator signals (slow-varying envelopes in the time domain) at each acoustic frequency subband [108]. A different approach taken in a realtime recognition system for music signals measures sparseness to determine similarity between the trained model and modulation spectrum of a test signal [52]. This sparseness similarity measurement is used as a constraint in a gradient descent algorithm for non-negative matrix decomposition. As for assessing overall results, the latest researchers in modulation filtering simply use log-likelihood ratio (LLR), SNR, WSS, and PESQ for overall assessment measures [108] [109].

With several measures in the time and frequency domains and a few modulation spectral domain measures, we focus our research on developing a dissimilarity measure solely for modulation spectra with the additional benefit of a weighting function for acoustic and modulation frequencies while encompassing valuable aspects of most of the aforementioned assessment techniques. The benefits for the WMSD measure include a close perceptual relationship to the human auditory system, ability to quantify quality of modulation spectral features, usefulness of weighting in multiple frequency types, invariance of noise and other artifacts which are caused by transforming between domains, and comparison of long-term modulation features in signals (especially since temporal characteristics are significant in determining similarity [110]). Also, by exploiting resolution in the modulation spectrum, the WMSD is advantageous in comparisons between mis-aligned time windows of signals (since time and frequency-versus-time comparison methods that require signals to be directly aligned in time) as long as the period is the same. This method is similar

to finding rhythmic similarity of music by a method involving dynamic periodicity warping in [79]. Since modulation envelopes are consistent across any time window segments covering the same periodicity of a signal, modulation spectra are invariant to slight misalignments in time windows between signals. Also, modulation spectra with appropriate parameters may appear almost identical for different signals. This property leads to WMSD measures that may become invariant to a number of characteristics, depending on the application.

A.2 Experimental Results

Experiments consisted of a variety a signals and goals. We used 30 unique audio signals of 5-10 seconds in length with a sampling rate of 16 kHz. These audio signals were taken from commonly used datasets (TIMIT, RWC, University of Iowa Musical Instrument Sounds Database, LabRosa, TRIOS source separation dataset). The signals consisted of sentences from multiple speakers, single musical instrument notes, harmonic song segments, percussive song segments, and mixtures (with and without vocals). The modulation spectrum was computed (using the Modulation Toolbox for MATLAB) for each signal using 20 Hz acoustic frequency subband widths and modulation frequency resolution varying from 0.25 Hz to 2 Hz. Also, we varied weights with $w = 0, 0.1, 0.5, 1$, and 2, where we learn that greater weights generally provide a greater range of dissimilarity.

A.2.1 Verification

We first conducted verification experiments to show that the WMSD measure exhibited the same or similar characteristic patterns overall as the SNR for various weights. Also, when adding random noise of increasing levels to a signal, the WMSD increases (see example in Fig. 43). This WMSD versus SNR pattern was tested with separate sources of multi-source signals (see example in Fig. 44). For various weight values, the characteristic plots of the WMSD were consistent in how they increased and decreased (see speech example in Fig. 45). We finally computed the mean and standard deviation of the WMSDs and

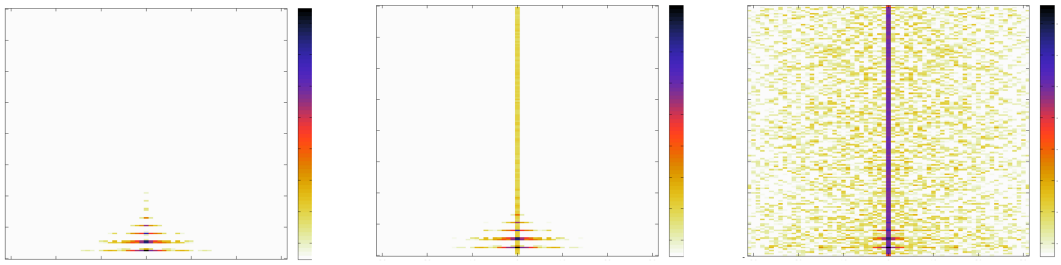


Figure 43: Modulation spectra shapes of a horn signal (middle-C note being repeating every second) with three levels of increasing noise added (from left to right).

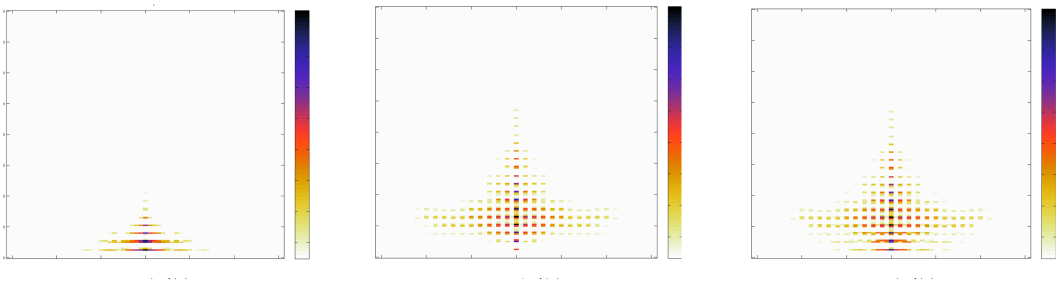


Figure 44: Modulation spectra shapes of a horn, trumpet, and combination of both (from left to right).

compared them with that of the SNR for all experiments. Table 9 shows that in all cases, the standard deviations were less than that of the SNRs, indicating that the WMSD measure is consistent overall.

A.2.2 Modulation Spectra Resolution vs. Tempo

We used various modulation frequency resolutions along the x-axis of the modulation spectrum: 0.25 Hz, 0.5 Hz, 1 Hz, and 2 Hz. An example signal (X_1) had an original tempo which

Table 9: Mean and standard deviations (in parentheses) of SNR (dB) and WMSD for all experiments.

	WMSD
$w = 0$	3.05(2.75)
$w = 0.1$	3.05(2.74)
$w = 0.5$	3.07(2.77)
$w = 1$	3.13(2.88)
$w = 2$	3.24(3.13)
SNR	-1.17e-16(5.81)

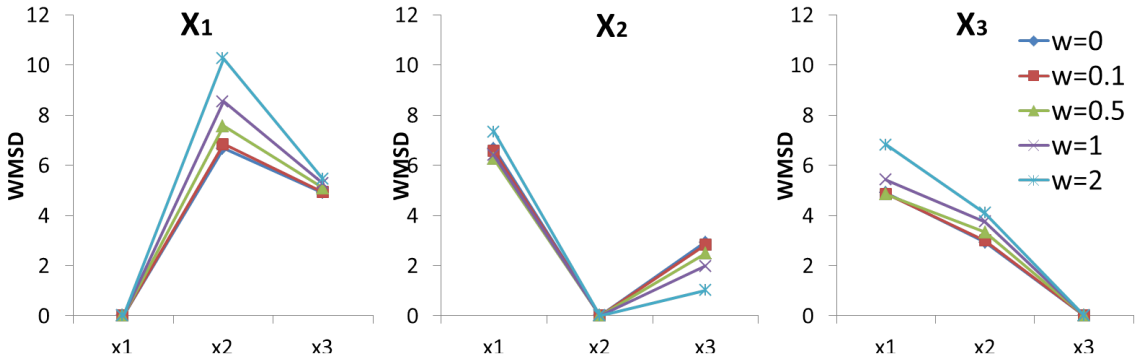


Figure 45: Characteristic plots with varying weights of WMSDs for X_1, X_2, and X_3 (from left to right) as it compares to each, where X_1 is a signal containing multiple speakers, X_2 is a percussive rhythmic signal with periodic horn and flute sounds, and X_3 contains both.

Table 10: Resolution vs. Tempo: Original (X_1), fast (X_2), and slow (X_3) tempo signals at fine (0.25 Hz) and broad (2 Hz) resolutions

Fine	WMSD(X_1,X_2)	WMSD(X_1,X_3)	WMSD(X_2,X_3)
$w = 0$	2.6419	2.6069	2.5592
$w = 2$	2.5851	2.7673	2.8381
Broad	WMSD(X_1,X_2)	WMSD(X_1,X_3)	WMSD(X_2,X_3)
$w = 0$	2.6019	2.5794	2.3640
$w = 2$	2.1007	2.0281	2.4185
SNR (db)	-0.0233	0.2425	0.0605

was decreased to 75% of its original tempo (result in signal X_2) and increased to 125% of its original tempo (result in signal X_3) without changing the pitch. We show the fine (0.25 Hz) and broad (2 Hz) resolutions of spectra in Fig. 46 and the WMSDs of the signals with varying weights and resolutions in Table 10. The WMSDs of the broad resolution are lesser than those of the higher resolution (0.25 Hz), which indicates greater similarity between the signals for lower modulation frequency resolutions. We also show that when all weights are equal to one as $w = 0$, i.e. with lesser emphasis on modulation and more focus on acoustics, signals X_2 and X_3 appear more similar than the other combinations. All of these differences are not as apparent with the SNR (since SNR and WMSD would ideally have an inverse relationship).

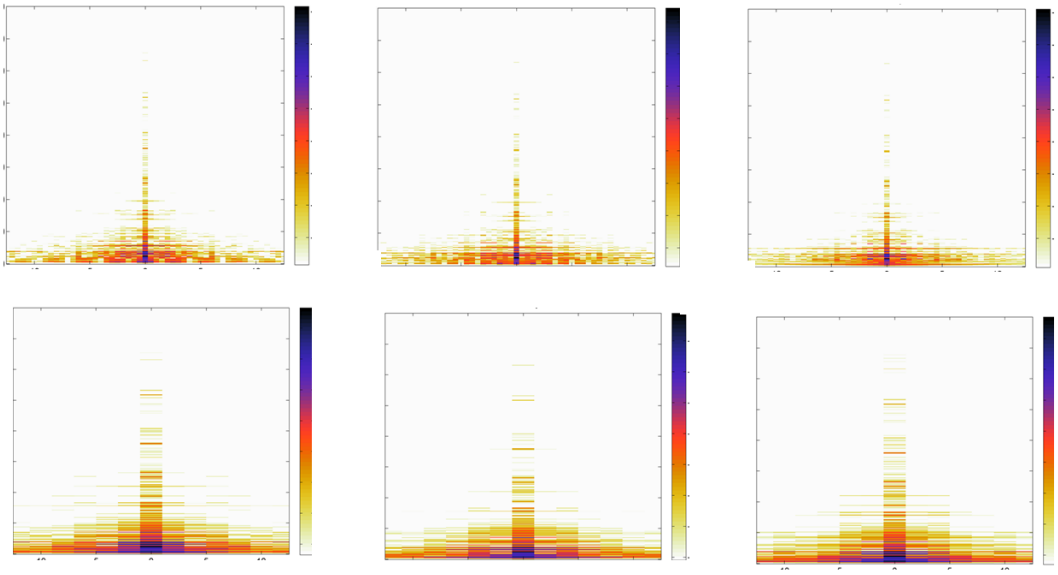


Figure 46: Modulation spectra shapes of same signals at original, fast, and slow tempos (from left to right) with fine resolution (top row) and broad resolution (bottom row).

A.2.3 Music Genre, Synthetic vs. Authentic, and Vibrato

Other experiments conducted included similarity between song segments from different genres that are somewhat subjective in comparison. Our results indicated that the weights may affect the subjective similarity. For one experiment comparing techno, pop, and swing-style jazz signals, with the exception of $w = 0$ and $w = 0.5$, all comparisons show that the jazz and pop signals were least similar, the techno and pop signals were moderately similar, and the techno and jazz signals were most similar. Another example experiment compared synthesized MIDI versions of song segments with live-recorded versions. The WMSD showed that although the compared versions may have different harmonics in acoustic frequency, similarity still existed amongst the modulation frequency features. Our final example experiment compared three, single middle-C trumpet notes: one without vibrato at a medium volume, one with vibrato at lower volume, and one with vibrato also at a medium volume. All notes are mostly similar in acoustic frequency harmonics while the vibrato notes are more similar in modulation frequency, especially with higher weight values ($w = 1, w = 2$). The experiments affirm that WMSD may also be used to compare

styles of multiple players, singers, or speakers.

A.3 Summary

WMSD may provide a useful evaluation method in the modulation spectral domain to measure accuracy, perceptual quality, and similarity of signals. Under certain conditions (i.e., parameters of the modulation spectrum, weighting of the acoustic and modulation frequencies, and scaling of the modulation spectrum's magnitudes), the WMSD measure may be applicable for comparing modulation features of signals regardless of noise, volume-level, time-window alignment, and tempo differences. Also, the WMSD measure may be beneficial for applications in gradient and iterative comparison methods, source separation algorithms with modulation filtering, and classification methods. Future work with the WMSD measure may consider *composite measures* [104] to better predict subjective quality by combining WMSD with other measures using multiple linear regression, thus finding a maximum subjective correlation. Future work may also include a more thorough study on how weights may be chosen depending on the signal type along with conducting a formal listening test to verify results with subjective evaluation data.

REFERENCES

[1] S. M. Schimmel, *Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices*. Thesis (phd), University of Washington, 2007.

[2] D. Byrd and M. Fingerhut, "The History of ISMIR- A Short Happy Tale," *D-Lib Magazine*, vol. 8, Nov. 2002.

[3] N. H. Sephus, A. D. Lanterman, and D. V. Anderson, "Modulation Spectral Features: In Pursuit of Invariant Representations of Music with Application to Unsupervised Source Identification," *Journal of New Music Research*, vol. To appear, no. Special Issue on Music Rhythm, 2014.

[4] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Acoustics, Speech, and Signal Processing, 1997 IEEE International Conference on*, vol. 3 of *ICASSP '97*, pp. 1647–1650 vol.3, Int. Comput. Sci. Inst., Berkeley, CA, IEEE, Apr. 1997.

[5] A. Eronen, "Comparison of features for musical instrument recognition," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pp. 19–22, Signal Process. Lab., Tampere Univ. of Technol., IEEE, 2001.

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[7] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Appl. Signal Process.*, vol. 2003, pp. 668–675, Jan. 2003.

[8] L. Atlas, P. Clark, and S. Schimmel, "Modulation Toolbox Version 2.1 for MATLAB." http://isdl.ee.washington.edu/projects/modulationtoolbox/, Sept. 2010.

[9] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using Independent Subspace Analysis," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, vol. 2003, pp. 843–848, 2003.

[10] H. D. Rohit, H. Deshp, and R. Singh, "Classification Of Music Signals In The Visual Domain," in *Proceedings of the COST G-6 Conference on Digital Audio Effects*, pp. DAFX1–DAFX4, 2001.

[11] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, S. Sagayama, and J. Le Roux, "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," in *16th European Signal Processing Conference. Proc. EUSIPCO'08.*, (Lausanne, Switzerland), The University of Tokyo, Aug. 2008.

[12] M.-J. Wu, Z.-S. Chen, J.-S. S. R. Jang, J.-M. Ren, Y.-H. Li, and C.-H. Lu, "Combining Visual and Acoustic Features for Music Genre Classification," in *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, vol. 2, pp. 124–129, Dept. of Comput. Sci., Nat. Tsing Hua Univ., Hsinchu, Taiwan, IEEE, Dec. 2011.

[13] N. Moritz, J. Anemüller, and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5492–5495, Project Group Hearing, Speech & Audio Technol., Fraunhofer IDMT, Oldenburg, Germany, IEEE, May 2011.

[14] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features," *Multimedia, IEEE Transactions on*, vol. 11, pp. 670–682, June 2009.

[15] A. Nagathil, P. Gottel, and R. Martin, "Hierarchical audio classification using cepstral modulation ratio regressions based on Legendre polynomials," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 2216–2219, IEEE, May 2011.

[16] V. Tyagi, I. McCowan, H. Misra, and H. Bourlard, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR," in *Automatic Speech Recognition and Understanding, 2003 IEEE Workshop on*, pp. 399–404, Dalle Molle Inst. for Perceptual Artificial Intelligence, Martigny, Switzerland, IEEE, Nov. 2003.

[17] C.-H. Lee, C.-H. Chou, C.-C. Lien, and J.-C. Fang, "Music genre classification using modulation spectral features and multiple prototype vectors representation," in *Image and Signal Processing (CISP), 2011 4th International Congress on*, vol. 5, pp. 2762–2766, Dept. of Comput. Sci. & Inf. Eng., Chung Hua Univ., Hsinchu, Taiwan, IEEE, Oct. 2011.

[18] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *Information Theory, IEEE Transactions on*, vol. 38, pp. 824–839, Mar. 1992.

[19] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.

[20] G. Langner, "Temporal processing of pitch in the auditory system," *Journal of New Music Research*, vol. 26, pp. 116–132, June 1997.

[21] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.

[22] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech.," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, Apr. 1990.

[23] S. Disch and B. Edler, "Multiband perceptual modulation analysis, processing and synthesis of audio signals," in *Acoustics, Speech and Signal Processing, 2009 IEEE International Conference on*, pp. 2305–2308, Lab. fur Informationstechnologie, Leibniz Univ. Hannover, Hannover, IEEE, Apr. 2009.

[24] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis.," *Trends in neurosciences*, vol. 34, pp. 114–123, Mar. 2011.

[25] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[26] J. H. McDermott and E. P. Simoncelli, "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis," *Neuron*, vol. 71, pp. 926–940, Sept. 2011.

[27] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[28] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," in *International Symposium on Music Information Retrieval*, 2000.

[29] M. Muller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal Processing for Music Analysis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, pp. 1088–1110, Oct. 2011.

[30] S. Ganapathy, S. Thomas, and H. Hermansky, "Modulation frequency features for phoneme recognition in noisy speech.," *The Journal of the Acoustical Society of America*, vol. 125, pp. EL8–EL12, Jan. 2009.

[31] J.-H. H. Bach, B. Kollmeier, and J. Anemuller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 41–44, Dept. of Phys., Carl von Ossietzky Univ. Oldenburg, Oldenburg, Germany, IEEE, Mar. 2010.

[32] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, Aug. 1998.

[33] L. Atlas, "Modulation Spectral Transforms: Application to Speech Separation and Modification," technical report, The Institute of Electronics, Information and Communication Engineers, Kyoto, Japan, June 2003.

[34] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, pp. 450–475, May 2010.

[35] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, pp. 282–305, Feb. 2012.

[36] Q. Li and L. Atlas, "Properties for modulation spectral filtering," in *Acoustics, Speech, and Signal Processing, 2005 IEEE International Conference on*, vol. 4, pp. iv/521–iv/524 Vol. 4, Dept. of Electr. Eng., Univ. of Washington, Seattle, WA, USA, IEEE, Mar. 2005.

[37] T. Kinnunen, "Joint Acoustic-Modulation Frequency for Speaker Recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on*, 2006.

[38] N. Malyska, T. F. Quatieri, and D. Sturim, "Automatic Dysphonia Recognition using Biologically-Inspired Amplitude-Modulation Features," in *Acoustics, Speech, and Signal Processing, 2005 IEEE International Conference on*, vol. 1, (Philadelphia, Pennsylvania, USA), pp. 873–876, IEEE, Mar. 2005.

[39] O. M. Mubarak, E. Ambikairajah, J. Epps, and T. S. Gunawan, "Modulation Features for Speech and Music Classification," in *Communication systems, 2006. ICCS 2006. 10th IEEE Singapore International Conference on*, pp. 1–5, Sch. of Electr. Eng. & Telecommun., New South Wales Univ., Sydney, NSW, IEEE, Oct. 2006.

[40] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5448–5451, Center for Robust Speech Syst. (CRSS), Univ. of Texas at Dallas, Richardson, TX, USA, IEEE, May 2011.

[41] S. Ganapathy, S. Thomas, and H. Hermansky, "Static and Dynamic Modulation Spectrum for Speech Recognition," *in Proc. of Interspeech 2009*, 2009.

[42] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from non-speech based on multiscale spectro-temporal Modulations," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 920–930, May 2006.

[43] J. Anemüller, D. Schmidt, and J.-H. Bach, "Detection of Speech Embedded in Real Acoustic Background Based on Amplitude Modulation Spectrogram Features," in *Proc. INTERSPEECH '08*, pp. 2582–2585, Interspeech, 2008.

[44] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-Negative Multilinear Principal Component Analysis of Auditory Temporal Modulations for Music Genre Classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 576–588, Mar. 2010.

[45] M. Markaki and Y. Stylianou, "Discrimination of speech from nonspeeech in broadcast news based on modulation frequency features," *Speech Communication*, vol. 53, pp. 726–735, May 2011.

[46] J. H. Bach, J. Anemüller, and B. Kollmeier, "Robust speech detection in real acoustic backgrounds with perceptually motivated features," *Speech Communication*, vol. 53, pp. 690–706, May 2011.

[47] T. H. Falk, F. J. Fraga, L. Trambaiolli, and R. Anghinah, "EEG Amplitude Modulation Analysis for Semi-Automated Diagnosis of Alzheimer's Disease," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 1–9, Aug. 2012.

[48] J. P. Havlicek, A. C. Bovik, and P. Maragos, "Modulation models for image processing and wavelet-based image demodulation," in *Signals, Systems and Computers, 1992 Conference Record of The Twenty-Sixth Asilomar Conference on*, pp. 805–810 vol.2, Lab. for Vision Syst., Texas Univ., Austin, TX, IEEE, Oct. 1992.

[49] N. Delprat, "Global frequency modulation laws extraction from the Gabor transform of a signal: a first study of the interacting components case," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, pp. 64–71, Jan. 1997.

[50] A. Zlatintsi and P. Maragos, "AM-FM Modulation Features for Music Instrument Signal Analysis and Recognition," in *20th European Signal Processing Conference Proc. EUSIPCO'12.*, EUSIPCO, 2012.

[51] M. Markaki and Y. Stylianou, "Evaluation of Modulation Frequency Features For Speaker Verification And Identification," *17th European Signal Processing Conference*, Aug. 2009.

[52] A. Cont, S. Dubnov, and D. Wessel, "Realtime Multiple-Pitch and Multiple-Instrument Recognition for Music Signals Using Sparse Non-Negative Constraints," in *10th International Conference on Digital Audio Effects (DAFx-07)*, pp. 85–92, 2007.

[53] J. Kauppinen, "Music Data Mining edited by Tao Li, Mitsunori Ogihara, George Tzanetakis," *International Statistical Review*, vol. 80, no. 1, pp. 189–190, 2012.

[54] B. D. Loeffler, "Instrument Timbres and Pitch Estimation in Polyphonic Music," thesis (master's), Georgia Institute of Technology, May 2006.

[55] F. J. Rodriguez-Serrano, P. Vera-Candeas, P. C. Molero, J. J. Carabias-Orti, and N. R. Reyes, "Amplitude Modulated Sinusoidal Modeling for Audio Onset Detection," in *18th European Signal Processing Conference*, pp. 512–516, EUSIPCO, Aug. 2010.

[56] L. M. Smith and H. Honing, "Time-Frequency Representation of Musical Rhythm by Continuous Wavelets," *Journal of Mathematics and Music*, vol. 2, pp. 81–97, Aug. 2008.

[57] M. Triki and D. T. M. Slock, "Periodic signal extraction with global amplitude and phase modulation for music signal decomposition," in *Acoustics, Speech, and Signal Processing, 2005 IEEE International Conference on*, vol. 3, pp. iii/233–iii/236 Vol. 3, Eurecom Inst., Sophia Antipolis, France, IEEE, Mar. 2005.

[58] C.-H. Lee, J.-L. Shih, K.-M. Yu, and J.-M. Su, "Automatic Music Genre Classification using Modulation Spectral Contrast Feature," in *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 204–207, Chung Hua Univ., Hsinchu, IEEE, July 2007.

[59] C.-H. Lee, H.-S. Lin, C.-H. Chou, and J.-L. Shih, "Modulation Spectral Analysis of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification," in *Intelligent Information Hiding and Multimedia Signal Processing, Fifth International Conference on*, pp. 1030–1033, Dept. of Comput. Sci. & Inf. Eng., Chung Hua Univ., Hsinchu, Taiwan, IEEE, 2009.

[60] A. Nagathil, T. Gerkmann, and R. Martin, "Musical Genre Classification Based on a Highly-Resolved Cepstral Modulation Spectrum," in *18th European Signal Processing Conference*, EUSIPCO, 2010.

[61] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music Genre Classification via Sparse Representations of Auditory Temporal Modulations," in *17th European Signal Processing Conference Proc.*, (Glasgow, Scotland), Aug. 2009.

[62] Y.-Y. Shi, X. Zhu, H.-G. Kim, and K.-W. Eom, "A Tempo Feature via Modulation Spectrum Analysis and its Application to Music Emotion Classification," in *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 1085–1088, IEEE, July 2006.

[63] S.-C. Lim, S.-J. Jang, S.-P. Lee, and M. Y. Kim, "Music genre/mood classification using a feature-based modulation spectrum," in *Mobile IT Convergence (ICMIC), 2011 International Conference on*, pp. 133–136, Dept. Inf. Commun. Eng., Sejong Univ., Seoul, South Korea, IEEE, 2011.

[64] K. Jensen, *Timbre models of musical sounds*. Thesis (phd), University of Copenhagen, 1999.

[65] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. A Bradford Book, Sept. 1994.

[66] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *The Journal of the Acoustical Society of America*, vol. 61, pp. 1270–1277, May 1977.

[67] V. Alluri and P. Toiviainen, "Exploring Perceptual and Acoustical Correlates of Polyphonic Timbre," *Music Perception*, vol. 27, no. 3, pp. 223–241, 2009.

[68] P. Ru and S. A. Shamma, "Representation of musical timbre in the auditory cortex," *Journal of New Music Research*, vol. 26, pp. 154–169, June 1997.

[69] C. Joder, S. Essid, and G. Richard, "Temporal Integration for Audio Classification With Application to Musical Instrument Classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, pp. 174–186, Jan. 2009.

[70] P. Herrera, G. Peeters, and S. Dubnov, "Automatic Classification of Musical Instrument Sounds," *Journal of New Music Research*, vol. 32, no. 1, pp. 3–21, 2003.

[71] S. K. Kopparapu, M. A. Pandharipande, and G. Sita, "Music and vocal separation using multiband modulation based features," in *Industrial Electronics and; Applications (ISIEA), 2010 IEEE Symposium on*, pp. 733–737, TCS Innovation Lab., Tata Consultancy Services Ltd., Mumbai, India, IEEE, Oct. 2010.

[72] E. Tsunoo, N. Ono, and S. Sagayama, "Rhythm map: Extraction of unit rhythmic patterns and analysis of rhythmic structure from music acoustic signals," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, (Taipei, Taiwan), pp. 185–188, IEEE, Apr. 2009.

[73] L. Atlas and C. Janssen, "Coherent modulation spectral filtering for single-channel music source separation," in *Acoustics, Speech, and Signal Processing, 2005 IEEE International Conference on*, vol. 4, pp. iv/461–iv/464 Vol. 4, Dept. of Electr. Eng., Univ. of Washington, Seattle, WA, USA, IEEE, Mar. 2005.

[74] Y. Li, J. Woodruff, and D. Wang, "Monaural Musical Sound Separation Based on Pitch and Common Amplitude Modulation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1361–1371, 2009.

[75] D. Barry, D. Fitzgerald, E. Coyle and B. Lawlor, "Drum Source Separation using Percussive Feature Detection and Spectral Modulation," in *IEE Irish Signals and Systems Conference*, pp. 13–17, 2005.

[76] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, pp. 240–254, May 2000.

[77] P. Clark and L. Atlas, "Time-Frequency Coherent Modulation Filtering of Nonstationary Signals," *IEEE Transactions on Signal Processing*, vol. 57, pp. 4323–4332, Nov. 2009.

[78] J.-W. Suh, S. O. Sadjadi, G. Liu, T. Hasan, K. W. Godin, and J. H. L. Hansen, "Exploring Hilbert envelope based acoustic features in i-vector speaker verication using HT-PLDA," *Proc. of NIST 2011 Speaker Recognition Evaluation Workshop*, 2011.

[79] A. Holzapfel and Y. Stylianou, "Rhythmic similarity of music based on dynamic periodicity warping," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2217–2220, IEEE, Mar. 2008.

[80] S. M. Schimmel, "Analysis of signal reconstruction after modulation filtering," in *Proc. SPIE*, vol. 5910, pp. 59100H–59100H–10, 2005.

[81] N. Sephus, A. Lanterman, and D. Anderson, "Exploring Frequency Modulation Features and Resolution in the Modulation Spectrum," in *2013 IEEE Digital Signal Processing (DSP) and Signal Processing Education (SPE) Meeting*, (Napa, CA), pp. 169–174, 2013.

[82] J. Chowning and D. Bristow, *FM Theory and Applications: By Musicians for Musicians*. Hal Leonard Corp, 1986.

[83] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proceedings of the 7th international conference on Independent component analysis and signal separation*, ICA'07, (London, UK), pp. 414–421, Springer-Verlag, 2007.

[84] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Proceedings of the 9th international conference on Latent variable analysis and signal separation*, LVA/ICA'10, (St. Malo, France), pp. 140–148, Springer-Verlag, 2010.

[85] K. Yoshii and M. Goto, "Infinite Composite Autoregressive Models for Music Signal Analysis," in *13th International Society of Music Information Retrieval (ISMIR'12)*, pp. 79–84, 2012.

[86] O. L. Smart, N. H. Sephus, and R. E. Gross, "Application of Modulation Spectrum for iEEG Seizure Analysis," in *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, pp. Recently accepted–To be published in May 2014 proc, 2014.

[87] R. S. Fisher, W. van Emde Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel J., "Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, 2005.

[88] F. Rosenow and H. Luders, "Presurgical evaluation of epilepsy," *Brain*, vol. 124, no. Pt 9, pp. 1683–1700, 2001.

[89] G. A. James, S. P. Tripathi, J. G. Ojemann, R. E. Gross, and D. L. Drane, "Diminished default mode network recruitment of the hippocampus and parahippocampus in temporal lobe epilepsy," *J Neurosurg*, vol. 119, no. 2, pp. 288–300, 2013.

[90] A. Schnitzler and J. Gross, "Normal and pathological oscillatory communication in the brain," *Nat Rev Neurosci*, vol. 6, no. 4, pp. 285–296, 2005.

[91] J. E. Lisman and O. Jensen, "The theta-gamma neural code," *Neuron*, vol. 77, no. 6, pp. 1002–1016, 2013.

[92] G. Buzsaki and X. J. Wang, "Mechanisms of gamma oscillations," *Annu Rev Neurosci*, vol. 35, pp. 203–225, 2012.

[93] A. Ivanov and X. Chen, "Modulation Spectrum Analysis for Speaker Personality Trait Recognition," in *INTERSPEECH*, ISCA.

[94] M. Guirgis, Y. Chinvarun, P. L. Carlen, and B. L. Bardakjian, "The role of delta-modulated high frequency oscillations in seizure state classification," *Conf Proc IEEE Eng Med Biol Soc*, vol. 2013, pp. 6595–6598, 2013.

[95] T. H. Falk and C. Wai-Yip, "Modulation filtering for heart and lung sound separation from breath sound recordings," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 1859–1862, Aug. 2008.

[96] S. A. Weiss, G. P. Banks, J. McKhann G. M., R. R. Goodman, R. G. Emerson, A. J. Trevelyan, and C. A. Schevon, "Ictal high frequency oscillations distinguish two types of seizure territories in humans," *Brain*, 2013.

[97] G. M. Ibrahim, R. Anderson, T. Akiyama, A. Ochi, H. Otsubo, G. Singh-Cadieux, E. Donner, J. T. Rutka, r. Snead O. C., and S. M. Doesburg, "Neocortical pathological high-frequency oscillations are associated with frequency-dependent alterations in functional network topology," *J Neurophysiol*, vol. 110, no. 10, pp. 2475–2483, 2013.

[98] C. Ramon and M. D. Holmes, "Stochastic Behavior of Phase Synchronization Index and Cross-Frequency Couplings in Epileptogenic Zones during Interictal Periods Measured with Scalp dEEG," *Front Neurol*, vol. 4, p. 57, 2013.

[99] R. T. Canolty, E. Edwards, S. S. Dalal, M. Soltani, S. S. Nagarajan, H. E. Kirsch, M. S. Berger, N. M. Barbaro, and R. T. Knight, "High gamma power is phase-locked to theta oscillations in human neocortex," *Science*, vol. 313, no. 5793, pp. 1626–1628, 2006.

[100] A. B. Tort, R. Komorowski, H. Eichenbaum, and N. Kopell, "Measuring phase-amplitude coupling between neuronal oscillations of different frequencies," *J Neurophysiol*, vol. 104, no. 2, pp. 1195–210, 2010. Tort, Adriano B L Komorowski, Robert Eichenbaum, Howard Kopell, Nancy MH-51570/MH/NIMH NIH HHS/United States MH-71702/MH/NIMH NIH HHS/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United States Journal of neurophysiology J Neurophysiol. 2010 Aug;104(2):1195-210. doi: 10.1152/jn.00106.2010. Epub 2010 May 12.

[101] M. A. Kramer and U. T. Eden, "Assessment of cross-frequency coupling with confidence using generalized linear models," *J Neurosci Methods*, vol. 220, no. 1, pp. 64–74, 2013.

[102] O. Smart, "Unsupervised seizure detection using modulation spectra measures: a preliminary study," 2014.

[103] N. H. Sephus, T. O. Olubanjo, and D. V. Anderson, "Enhancing Online Music Lessons with Applications in Automating Self-Learning Tutorials and Performance Assessment," in *12th IEEE International Conference on Machine Learning and Applications*, (Miami, FL), 2013.

[104] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. 2000.

[105] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 229–238, Jan. 2008.

[106] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Evaluating the Generalization of the Hearing Aid Speech Quality Index (HASQI)," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 407–415, Feb. 2013.

[107] A. Berenzweig, B. Logan, D. P. Ellis, and B. Whitman, "A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measures," *Computer Music Journal*, vol. 28, pp. 63–76, June 2004.

[108] A. Mahmoodzadeh, H. Sheikhzadeh, H. R. Abutalebi, and H. Soltanian-Zadeh, "A hybrid coherent-incoherent method of modulation filtering for Single Channel Speech Separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 329–332, Mar. 2012.

[109] K. Siedenburg and P. Depalle, "Modulation Filtering For Structured Time-Frequency Estimation of Audio Signals," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (New Paltz, NY), 2013.

[110] J. Reed and C.-H. Lee, "On the importance of modeling temporal information in music tag annotation," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1873–1876, IEEE, Apr. 2009.